Pararetrovirus-like sequences in the genome of plants

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Celia Napier Hansen
Department of Biology
University of Leicester

November 2002

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration

I hereby declare that no part of this thesis has been previously submitted to this or any other university as part of the requirements for a higher degree. The content of this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where it is acknowledged.

The work was conducted in the department of Biology, University of Leicester and John Innes Centre, Norwich during the period October 1999 to October 2002.

Signed.....Celia Hansen.....

Celia N. Hansen, November 2002

Pararetrovirus-like sequences in the genome of plants

by

Celia Napier Hansen

Abstract

Retroelements have been increasingly recognized as a major component of plant and animal genomes, and are characterized by the presence of genes for the enzyme reverse transcriptase. Pararetroviruses are well known in plants as episomal, infective DNA particles containing reverse transcriptase genes. Recent research has presented evidence that pararetroviruses are also integrated into the nuclear genome of certain plant families. The work of this thesis indicates that this might be a more widespread phenomenon.

Primers were designed to reveal pararetrovirus-like sequences from potato (*Solanum tuberosum*) by PCR. One primer pair covered an area of the reverse transcriptase region, others covered the end of a transactivator and a repetitive region. The sequences obtained ranged from 800 to 1100 bp and had homology to known sequences of TPV (tobacco pararetrovirus) and TVCV (*Tobacco vein clearing virus*). Aligning the sequences from the potato genome gave a phylogenetic clustering into three different groups. Genomic Southern hybridization was used to confirm the results, with probes hybridizing to the full length of the digested DNA. *In situ* hybridization was used to localize different pararetrovirus-like sequences to the chromosomes of potato, which showed a low copy number sequence dispersed throughout the genome.

A diverse range of plant species was analysed with the pararetrovirus-like primers, and clones homologous to TPV and TVCV were obtained from a liverwort, a fern, tomato, tobacco, pea, rice and banana. These results were confirmed with genomic Southern hybridization, which also revealed pararetrovirus-like sequences in the DNA of other lower plants, Gymnosperms and Angiosperms.

Integrated pararetroviruses have been found to be a serious problem in banana breeding as micro-propagation may give rise to episomal viruses, reducing the yield considerably. The possibility of the widespread presence of pararetrovirus-like sequences in plants may have important implications for virus resistance, sudden pathogen outbreak and genome evolution.

Acknowledgements

I would like to thank the following people:

First of all Dr Glyn Harper (John Innes, Norwich) for always being encouraging, helpful and positive. He "housed" me during the period of moving to Leicester, did some of the radioactive work for me and helped generating phylogenetic trees. And not least for reading the thesis draft and giving useful critics and comments.

Dr Nikolai Friesen (previous John Innes, Norwich) for helping me in the difficult start period and introducing me to and helping with molecular techniques.

Dr John Bailey (University of Leicester) for teaching me the technical details of his speciality - cytology. For reading and commenting upon part of the thesis and correcting my English - not always successfully. Also for arranging botanical excursions and being a friend.

Prof. Pat Heslop-Harrison for helping me with the greater outlines of experiments and the building up of the thesis and for doing some of the radioactive work.

Sybille Kubis and other people in the lab for useful help and cheerful company.

Abbreviations

bp	base pair
BSV	*Banana streak virus*
CaMV	*Cauliflower mosaic virus*
CsVMV	*Cassava vein mosaic virus*
DAPI	4',6-diamidino-2-phenylindole
ERV	endogenous retrovirus
EST	expressed sequence tags
FITC	fluorescein isothiocyanate
HERV	human endogenous retrovirus
ICTV	International Committee on the Taxonomy of Viruses
kb	kilo base
LINE	long interspersed nuclear element
LTR	long terminal repeat
NOR	nucleolar organizer region
nt	nucleotide
ORF	open reading frame
PCR	polymerase chain reaction
PRV-L	pararetrovirus-like
PTGS	post transcriptional gene silencing
PVCV	*Petunia vein clearing virus*
rDNA	ribosomal DNA
rt	room temperature
RT	reverse transcriptase
RTBV	*Rice tungro bacilliform virus*
SINE	short interspersed nuclear element
STC	sequence-tagged connectors
TAV	trans-activator
TPV	Tobacco pararetrovirus
TVCV	*Tobacco vein clearing virus*

For retroelement components see also table 1.1.

Table of content

Chapter I: Introduction: Retroelements and their interaction with the plant and animal genome

1.1 Transposable elements

Transposable elements are an important component of the genomes of all species of bacteria, animals, fungi and plants. In many of these species, the various elements or recognizable derivatives of them represent a significant fraction of the DNA, often being dispersed over much of the genome. Complete elements are typically 2 to 15 kb in length. The transposable elements can be divided into two major types: The DNA transposable elements which have the capacity to excise themselves and reintegrate elsewhere in the genome and the retroelements which replicate via an RNA intermediate, leaving the original element and forming a new one. These replicative properties mean that transposable elements are one of the most dynamic and rapidly evolving components of the genome.

1.1.1 DNA transposable elements

The presence of DNA transposable elements was first recognized by McClintock more than 50 years ago, as a result of their giving rise to phenotypic mutations in maize (McClintock, 1984). This ability to mutate the genome has been used extensively as a tool to isolate and characterize many structural and regulatory genes both in their original host and introduced in new plant species (McClintock, 1984; Wessler, 2001; Azpiroz-Leehan and Feldmann, 1997).

DNA transposable elements are bordered by terminal inverted repeats which are required for transposition. Autonomous elements encode a functional transposase and are able to mobilize and transpose themselves and also non-autonomous (mutated) or truncated elements (reviewed by Kunze *et al.*, 1997). DNA transposable elements are widespread in insect, plant and animal genomes and many different elements have been described. Among these the *mariner* element (Robertson, 1993; Auge-Gouillou *et al.*, 1995; Feschotte and Wessler, 2002) and the *P* element in *Drosophila* species (Daniels *et al.*, 1990) are most prominent. In plants the majority of elements have been grouped into *Ac*/*Ds* (activator/dissociation), *En/Spm* (enhancer/suppressor-mutator) and *Mu* (mutator) (for review Kunze *et al.*, 1997) and lately *En/Spm* elements have been identified and characterized in a range of plant species (Staginnus *et al.*, 2001; Kubis *et al.*, 2002). A thorough investigation of the

Arabidopsis genome, in course of the *Arabidopsis* genome sequencing project, found 623 transposable elements of which the DNA transposons accounted for 50% of the genome compared to 22% for the retroelements, the remaining 28% being unknown classes of transposable elements. Most of the transposons were found in intergenic regions (Le *et al.*, 2000).

1.1.2 Retroelements

Retroelements have a close relationship to particular orders of viruses through, among other, the shared possession of the gene for reverse transcriptase (RT). RT (RNA-directed DNA polymerase, EC 2.2.7.49) is capable of using an RNA template to make a complementary DNA molecule, enabling the element to convert between RNA and DNA form through the replication cycle. Xiong and Eickbush (1990) recognize the four most abundant groups of retroelements as: retroviruses, long terminal repeat (LTR)- and non-LTR retrotransposons and group II mitochondrial introns. The pararetroviruses, the major topic of this thesis and one which has had only limited previous study form an additional group. A virus can be defined as a minute, intracellular obligate parasite, visible only under the electron microscope. It is unable to multiply or express its genes outside of the host cell, as it requires host cell enzymes to aid DNA replication, transcription and translation (Lawrence, 1988). Retrotransposons were recognized early on as being a major evolutionary force in the host genome and some of the first elements were described in yeast and *Drosophila* (e.g. Clare and Farabaugh, 1985; Saigo *et al.*, 1984) and have often been the subject of reviews (Finnegan, 1985; Boeke and Corces, 1989; Kumar and Bennetzen, 1999). Retrotransposons can take up 50% or more of their host's nuclear genome (Kumar and Bennetzen, 1999; International Human Genome Sequencing Consortium, 2001) and retroelements are widespread in very diverse organisms from bacteria and yeasts to plants and animals. Retroelements may have no effect on the host or give severe illnesses; examples of the latter are the retrovirus Human Immunodeficiency Virus (HIV) and the pararetrovirus Hepatitis B (HBV). No retrovirus *senso stricto* has yet been found in plants although certain characteristic genes (*envelope*) have been identified in some *gypsy*-like elements (Petropoulos, 1997; Wright and Voytas, 2002).

The definition of retroelements has been diverse and in recent years two schools are starting to merge: the study of viruses and the study of repetitive sequences. Hull (1999), from the virus side, proposed a classification for reverse transcribing elements which includes viruses and the transposable elements with RT, also called classII elements or LTR and non-LTR elements. Hull (1999) called all elements that include RT retroelements, a usage which will be adopted here. Hull (1999) used the term retrotransposon for the LTR elements (*copia*-like and *gypsy*-like) and they are classified in the virus system as they, or at least *gypsy*, can form virus-like particles (International committee on taxonomy of viruses (ICTV), http://www.ncbi.nlm.nih.gov/ICTV/). In general terms the non-LTR LINE and SINE elements are often included in retrotransposons and they will also be included here, as non-LTR retrotransposons. The classification of the various retroelements will be described further below.

1.2 Classification of retroelements

If there is a single early origin of RT, then a natural classification of the retroelements is possible, with an evolutionary or phylogenetic foundation. Even if not, the division of genomes or genome segments containing RT provides a convenient framework for a classification. After many decades of flux, viral taxonomy is becoming stable with many divisions having a natural basis (Hull, 1999; Hull, 2002; ICTV). The viruses are classified using physical and biological properties including their biochemical composition, host range, replication strategy, particle structure and general genome organization. Though the genome and sequences of viruses has now become the dominant property used in classification which has grown to include other retroelements (Hull 2002; ICTV). Figure 1.1 presents a possible and simple model of relationships between RT containing elements. Retroelements can be divided into non-viral and viral forms as in figure 1.1, suggested by Hull (1999). Figure 1.2 shows an alternative scheme for phylogeny and thereby classification of retroelements presented by Xiong and Eickbush (1990) based on sequence data from the RT region. It would not be fair to say that one is more 'correct' than the other: both have a basic bipartite system. One is built generally on physical properties, the other builds on the amino acid differences and similarities in the RT region. The differences of these two systems will be discussed below (paragraph 1.4.3).

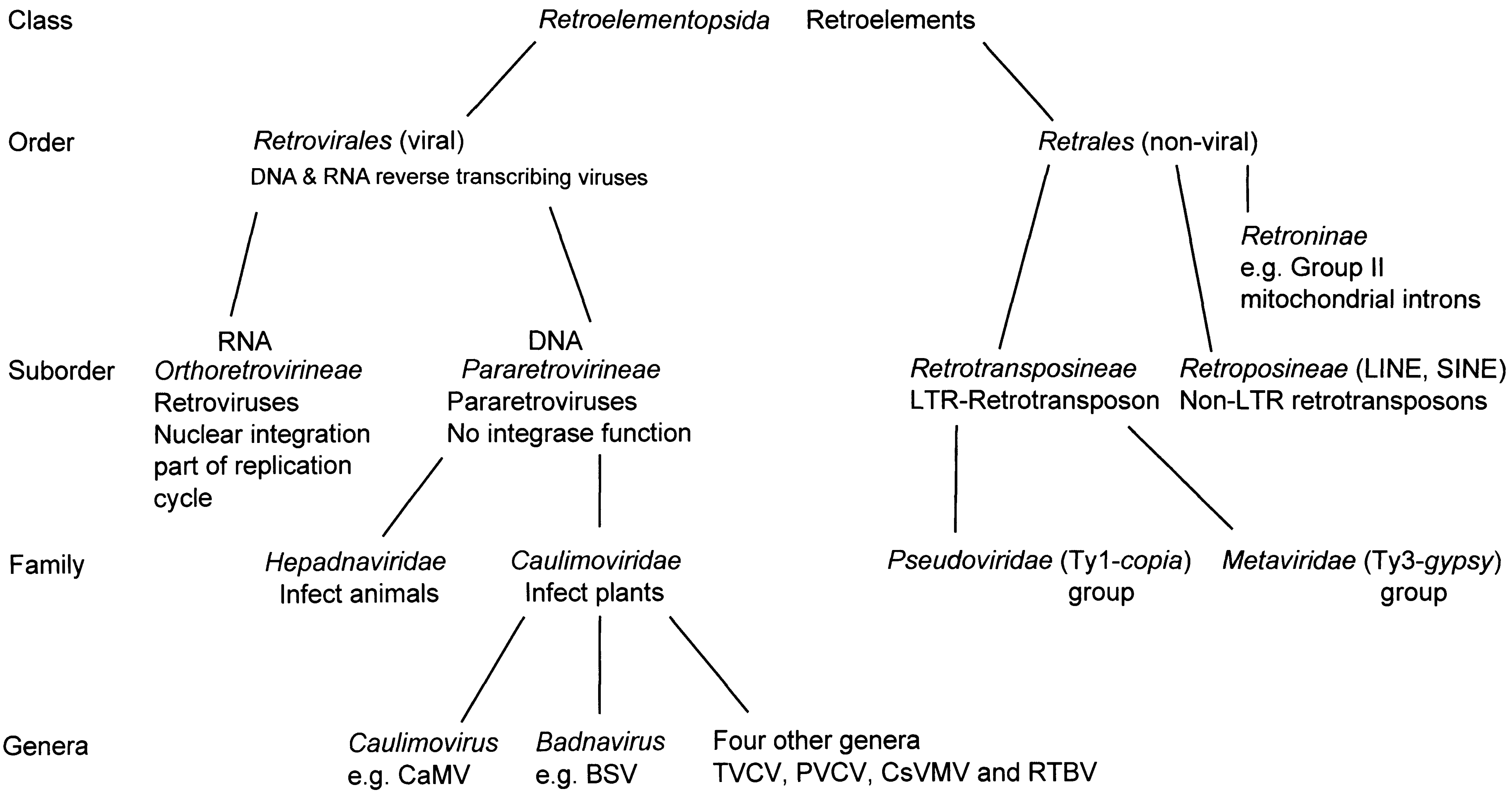


Figure 1.1. Diagram after Hull (1999) showing a classification of various retroelements. Abbreviations: CaMV, *Cauliflower mosaic virus*; BSV, *Banana streak virus*; TVCV, *Tobacco vein clearing virus*; PVCV, *Petunia vein clearing virus*; CsVMV, *Cassava vein mosaic virus*; RTBV, *Rice tungro bacilliform virus*. See figure 1.4 for examples of *copia*, *gypsy* and retrovirus sequences.

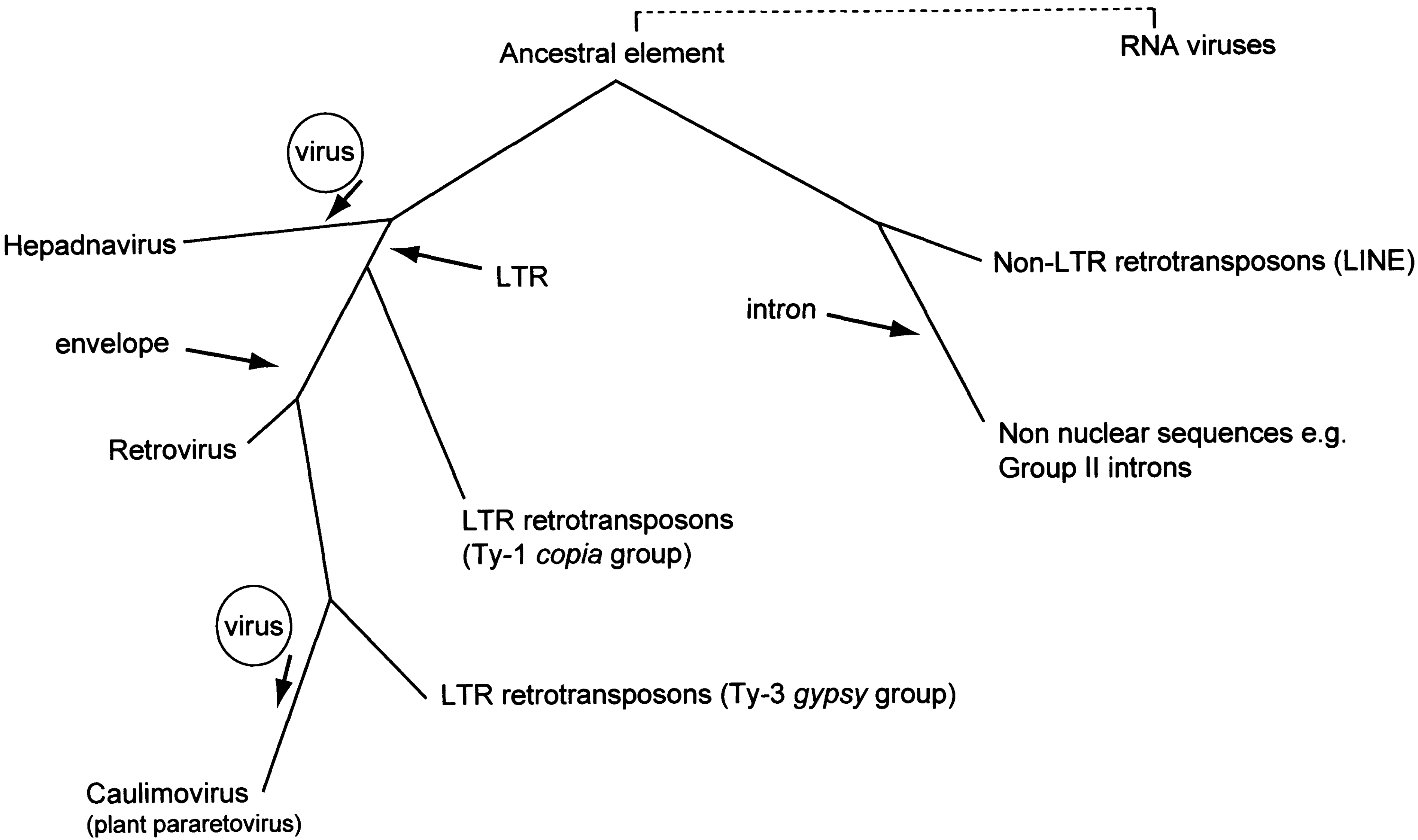

Ancestral element
RNA viruses
virus
Hepadnavirus
LTR
Non-LTR retrotransposons (LINE)
intron
envelope
Retrovirus
Non nuclear sequences e.g. Group II introns
LTR retrotransposons (Ty-1 *copia* group)
virus
LTR retrotransposons (Ty-3 *gypsy* group)
Caulimovirus
(plant pararetovirus)

Figure 1.2. Diagram showing the relationship and origin of retroelements based on alignment of the RT region, after Xiong and Eickbush (1990).

1.2.1 Non-viral retroelements

The order *Retrales* (Hull, 1999) holds the non-viral elements and includes the *Pseudoviridae* (*Ty1-copia* group) and the *Metaviridae* (*Ty3-gypsy* group), which have no known viral infectivity but can form virus-like particles (figure 1.1). A suborder *Retroposinae* (Hull, 1999) includes the LINEs and the truncated derivative SINEs. Within the *Retrales*, recent data show that the retrotransposons carrying LTRs (*copia* and *gypsy*-like) arose in eukaryotic organisms well after the non-LTR retroelements (LINE and SINE) (Malik and Eickbush, 2001). Sequences related to the *Retrales* are widespread in plants, fungi and insects as well as vertebrates such as lampreys, bony fish, amphibians, reptiles and mammals, although some groups, such as *copia*-like elements in mammals, may be missing in a few taxa (*copia*: Flavell *et al.*, 1992a; Flavell *et al.*, 1992b; Brandes *et al.*, 1997; Heslop-Harrison *et al.*, 1997; Flavell and Smith, 1992. *Gypsy*: Miller *et al.*, 1999; Suoniemi *et al.*, 1998; Kumekawa *et al.*, 1999. LINE and SINE: Kubis *et al.*, 1998; Schmidt, 1999; Malik *et al.*, 1999). *Copia* and *gypsy* elements often have one or two open reading frames (ORFs) containing *gag* and *pol* genes bordered by LTRs, though sometimes a third ORF is present. Structurally *copia* and *gypsy* differ in the order of the genes (see table 1.1 for the structure of a gypsy-like element). LINEs are simpler structures than *copia* and *gypsy* elements but still contain many of the same properties e.g. *gag* and *pol* but have an endonuclease function instead of integrase.

Another abundant group of non-viral retroelements are the group II mitochondrial introns, perhaps closer to the LINE elements than to the LTR retrotransposons. They are believed to have originated in bacteria and have spread to mitochondria and chloroplasts (Dai and Zimmerly, 2002). Group II introns exist in ORF-less and ORF containing forms. ORF-less introns contain six domains in a self-splicing RNA structure, and have a size of about 600 bp (Dai and Zimmerly, 2002). ORF containing introns encode a RT related reading frame in one of the domains (domain 4) and besides has a domain for splicing activity giving a total length of 2 to 3 kb. Sequences of a gene from two individuals may differ only by the insertion of a group II intron (Dai and Zimmerly, 2002; Zimmerly *et al.*, 2001).

1.2.2 Viral retroelements

The group of "DNA and RNA reverse transcribing viruses" (Pringle, 1999) or retrovirales (Hull, 1999), consists of elements potentially capable of infection and includes the retroviruses (RNA genome) and the pararetroviruses (DNA genome), see figure 1.1. The vertebrate retroviruses, of the *Orthoretrovirinae*, have an RNA genome in the infective form but are usually transcribed into DNA with RT to be integrated into the nuclear genome of the host. *Pararetrovirinae*, found in both plant and animal kingdoms, encapsulate a dsDNA genome and replicate through an RNA intermediate; no integrase function is detected in their genome i.e. integration is not an obligatory part of their replication cycle (Hull and Covey, 1996). Hull (1999) recognises two families of pararetroviruses, the animal viruses of the *Hepadnaviridae* and the plant viruses of the *Caulimoviridae*, the latter including six genera; *Badnavirus*, *Caulimovirus*, *Tobacco vein clearing virus* (TVCV), *Rice tungro bacilliform virus* (RTBV), *Petunia vein clearing virus* (PVCV) and *Cassava vein mosaic virus* (CsVMV) (Pringle, 1999). Caulimoviruses have a genome of about 7-8000 bp containing two or more ORFs; in their compact genomes, the ORFs often overlap by a small number of base pairs, typically 1-50. Caulimoviruses are not very well related on a molecular basis, with sequence comparison somewhat at the limit of interpretation for phylogenetic relationships (de Kochko *et al.*, 1998).

A subfamily of the retroviruses, the *Spumavirinae* (foamy viruses), is perhaps an intermediate between retroviruses and pararetroviruses. These differ from retroviruses in a number of ways one of which is the lack of the *gag* gene cysteine-histidine repeat motif, which is also missing in a member of *Hepadnavirideae* (HBV) (Lecellier and Saïb, 2000).

1.2.3 Between retrotransposon and virus - the *envelope* gene

It is a matter of discussion whether retrotransposons evolved from viruses or viruses evolved from retrotransposons. The *envelope* gene is probably the key point giving an element the ability to move between cells and individuals and thereby be infective i.e. ability for horizontal transmission. Genes encoding envelope functions are very heterogeneous at the sequence level and difficult to identify by homology even between retroviruses (Chavanne *et al.*, 1998). But although the *envelope* gene is not well conserved in the primary sequence both viral and retrotransposon envelope proteins share some structural similarities. They are typically translated from spliced mRNAs and the primary product encodes a signal peptide and a transmembrane domain near the C terminus (Wright and Voytas, 1998; Wright and Voytas, 2002; Chavanne *et al.*, 1998; Vicient *et al.*, 2001).

It has been proposed (Malik *et al.*, 2000) that a non-viral ancestor to errantiviruses (*Drosophila* specific *gypsy* virus) acquired the *envelope* gene from a baculovirus (dsDNA insect virus) as the *envelope* gene from these two insect viruses was found to share a common ancestry. Further more, baculoviruses were found to harbour inserts of LTR retrotransposons, which could be a step in the acquisition of an *envelope* gene by the latter. There are at least eight cases of *envelope*(-like) gene acquisition in the broad group of retroelements: *Sire1* from the *copia* group; *Athila*, *Cyclops*, *Osvaldo*, *Cer*, *Tas*, and errantiviruses from the *gypsy* group; Vertebrate retroviruses and plant caulimoviruses acquired an *envelope* gene from unknown sources - caulimoviruses could be the fusion of an LTR-retrotransposable element with a plant virus (Malik *et al.*, 2000; Chavanne *et al.*, 1998).

A partial solution to gain more mobility could be for a transposable element to insert into a (DNA based) viral genome and piggyback to a new genome (Malik *et al.*, 2000). Another point of view is that transposable elements are remnants of infectious viruses which have lost most of the *envelope* gene. The envelope function is less useful in plants, as having cell walls is an obstacle to membrane-membrane fusions by which the virus enters a cell (Bennetzen, 2000). The problem of finding homology to unknown ORFs or sequences in retrotransposons could be that some fragment of host DNA, and not an *envelope* gene, has been captured during transposition - called transduction (Chavanne *et al.*, 1998; Wright and Voytas, 2002).

1.2.4 Replicative cycle of retroelements

In the replicative cycle of plant pararetroviruses, gene expression starts with transcription from an episomal supercoiled dsDNA mini chromosome in the nucleus (Olszewski *et al.*, 1982; Hull and Covey, 1996; figure 1.3). Here episomal is defined as an extrachromosomal element replicating independently of the chromosomes. The transcribed viral RNA moves to the cytoplasm and serves as template for both translation of the gene products and reverse transcription. RT generates a new circular dsDNA, which becomes packaged into naked virion particles (Pfeiffer and Hohn, 1983). The final virus particle moves between cells via plasmodesmata and between individual plants by insect transmission: *Caulimovirus* usually by aphids, *Badnavirus* usually by mealy bugs. For hepadnaviruses the movement between cells is mediated by direct exit and entry of cell membranes. Upon entrance into a new plant cell the virus disassembles and releases the circular DNA genome, which is targeted to the nucleus where discontinuities (gaps) in the strands are repaired to generate a new mini chromosome (Richards *et al.*, 1981; Hull and Covey, 1996; Hohn and Fütterer, 1997; Hull, 2002). *Copia*-like, *gypsy*-like and retroviruses have a somewhat different replication strategy to that of pararetroviruses, as integration is a normal part of their life cycle. Transcription is from an integrated double stranded DNA form (provirus) and as for pararetroviruses, the mRNA transcripts serve two functions, being translated into the proteins for replication and serving as template for replication. The replication is started by the binding of a host tRNA to the primer binding site (PBS) which gives the starting point for RT and eventually give rise to a circular single stranded DNA molecule. The complementary DNA strand is then generated and the new copy is integrated into the DNA of the same cell by an integrase, which creates nicks in the DNA stands (Feuerbach *et al.*, 1997; Kumar and Bennetzen, 1999). Retrovirus mRNA is packaged within a protein particle (the virus particle) which buds out of the cell and fuses into a new recipient cell where the reverse transcription and integration takes place (Hull and Covey, 1996). The replication cycle of LINE elements is less well characterized but differs from that of the LTR retrotransposons by the reverse transcription occurring at the site of insertion (Feng *et al.*, 1996; Sassaman *et al.*, 1997).

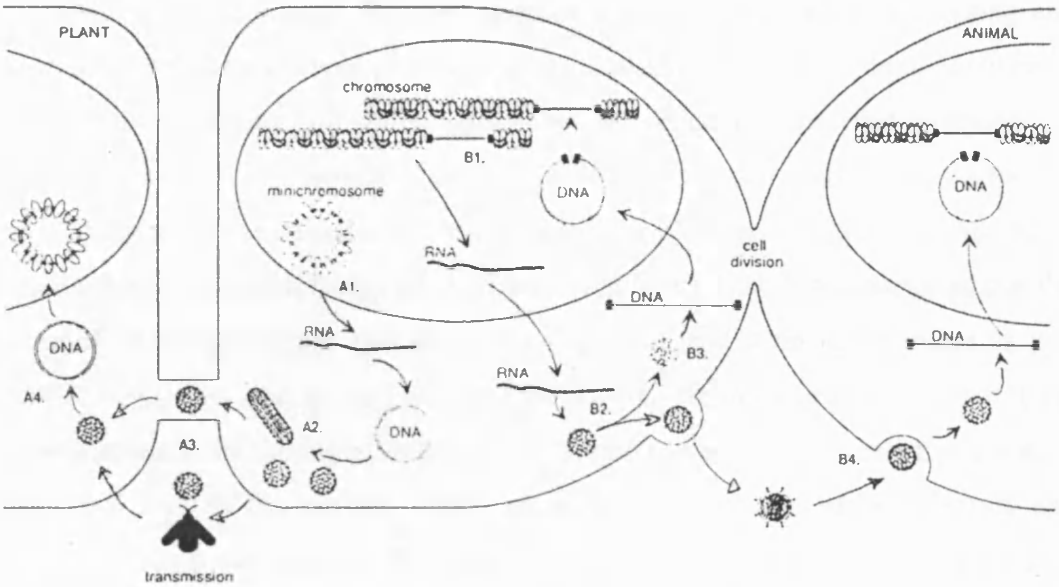


Figure 1.3. Replicative cycle of pararetroviruses, retrotransposons and retroviruses illustrated in a chimeric plant and animal cell (middle) and a plant (left) and animal (right) cell. Route A represent pararetrovirus replication from a minichromosome in the nucleus (A1), transcription and reverse transcription form of a circular dsDNA genome which is packaged (A2) and move to another plant cell via plasmodesmata or insect transmission (A3), the genome is released and migrate to the nucleus. Retrotransposons (route B1 to B3) integrated in the chromosomes are transcribed and reverse transcribed (B2-B3) to reintegrate at another chromosome location of the same cell. Retroviruses (route B1 to B4) get packaged in RNA form and bud out of the cell (B2) and into a new cell (B4) where the RNA genome is reverse transcribed and integrated in the chromosomes. After Hull and Covey (1996).

1.3 Origin of retroelements

The origin of retroelements, and subsequent evolution is highly speculative and is based on phylogenies of modern day isolations and sequences. It is estimated that the more ancient LINE elements date back at least 600 million years (Malik *et al*., 1999) and components of retroelements probably have an ancient origin going back with recognizable similarity to the RNA world. One of the reasons to believe this is that their RT is most similar to the RNA directed RNA polymerase of RNA viruses indicating that they share a common ancestor (Xiong and Eickbush, 1990). The RNA viruses are believed to be at least as old as retroelements as they have a greater diversity and are present in a wider variety of prokaryotes and eukaryotes than any retroelements. RT was discovered by Baltimore (1970) and Temin and Mizutani (1970) and is believed to be an ancient enzyme, which became necessary in the transition from RNA to DNA based sequences (Heslop-Harrison, 2000). RT is the only coding region which has complete conserved regions between all retroelements and is therefore often used for phylogenetic analysis of retroelements. Xiong and Eickbush (1990) found seven peptide regions (domains 1-7), which were common to all elements containing 178 amino acids. They identified 42 conserved positions that contained identical or chemically similar residues within the majority of 82 reverse transcribing sequences analysed. They rooted their phylogenetic tree with the RNA directed RNA polymerase from RNA viruses (see figure 1.2). It was assumed that the ancestral retrotransposable element had a *gag* gene and a *pol* gene either as two separate ORFs or one large ORF and no LTRs. Hepadnaviruses and non-LTR retrotransposons are the oldest branches on the tree supporting the general suggestion that LINE and SINEs are the oldest retroelements. For the hepadnaviruses and caulimoviruses it was assumed that only a portion of the *pol* gene, containing the RT-RNase H domain entered an already existing (virus)element. The retroviruses may represent a retroelement which acquired an *envelope* gene making it possible to leave the cell. For the retroelements of bacteria and organelles it was assumed that the RT region was captured by functional bacterial introns or organelle genomes or plasmids respectively (Xiong and Eickbush 1990).

Regardless of the method of analysis, it is still difficult to elucidate whether a function has been lost or gained and as more sequence information becomes available, evidence accumulate that genes encoded in the retroelement sequences have become exchanged between element classes during evolution (Lerat *et al.*, 1999; Malik and Eickbush, 2001). Elements missing or defective in some genes may exploit functions encoded in other elements belonging to the same or different classes that are present in the genome.

An unusual theory called viral eukaryogenesis proposes that the eukayotic nucleus evolved from a complex DNA virus (dsDNA *Poxviridae* type). It was proposed that the virus established a persistent presence in an archaeal cytoplasm and evolved into the leading nucleus by acquiring essential genes from the host genome. In support of this theory was stated some characteristic features of the eukaryotic nucleus which are not found in early bacteria or archaea such as mRNA capping, linear chromosomes, telomeres, separation of transcription from translation, and the sexual cycle (Bell, 2001).

1.4 Relationship between retroelements

Although the wide variety of retroelements belong to different orders and families, and reside in hosts throughout the living kingdoms they do have several conserved features in common. This is illustrated in figure 1.4, which shows selected retroelements drawn to relative scale, and discussed in the text below. The retroelements were chosen to represent the various types of elements, mostly from plants, or because homology was found in search for motifs. The information to make the drawings came from database sequences, several references and an extensive search for motifs. Figure 1.4 contrasts to virtually all other published figures of the structure of retroelements, which are said to be "not to scale". The retroelements in figure 1.4 are aligned through two completely conserved amino acid residues, aspartic acid (DD) in the RT region (Xiong and Eickbush, 1990). Table 1.1 lists abbreviations and functions for genes and other components of the retroelement genome. Most of the retroelements have their ORFs designated as *gag*, *pol* and eventually *envelope* whereas the plant viruses tend to have series from ORF1 to ORF3, 4 or above. The *gag* gene is an equivalent to the coat protein in viruses, and the *envelope* gene has an equivalent function to the movement protein of the plant pararetroviruses. The remarkable thing is that the arrangement of the different motifs is largely the same in all elements, the Cys-His motif always precedes the protease which is before the RT. The RNase H is located immediately after the RT. The integrase domain is situated after RNase H in *gypsy* elements and retroviruses, whereas integrase is between the protease and RT in *copia* elements. Pararetroviruses do not have an integrase domain. The *envelope* gene is situated as the last ORF in *gypsy* and retroviruses but in front of RT in pararetroviruses as a movement domain or function.

Legend to figure 1.4. Drawings of retroelements including a LINE, *copia* and *gypsy* elements, pararetroviruses and retroviruses. A scale in base pairs is shown at the bottom. The elements are manually aligned over the amino acids DD from RT. For the abbreviations of genes and other components see table 1.1. References are under the description of the individual elements paragraph 1.4.1.

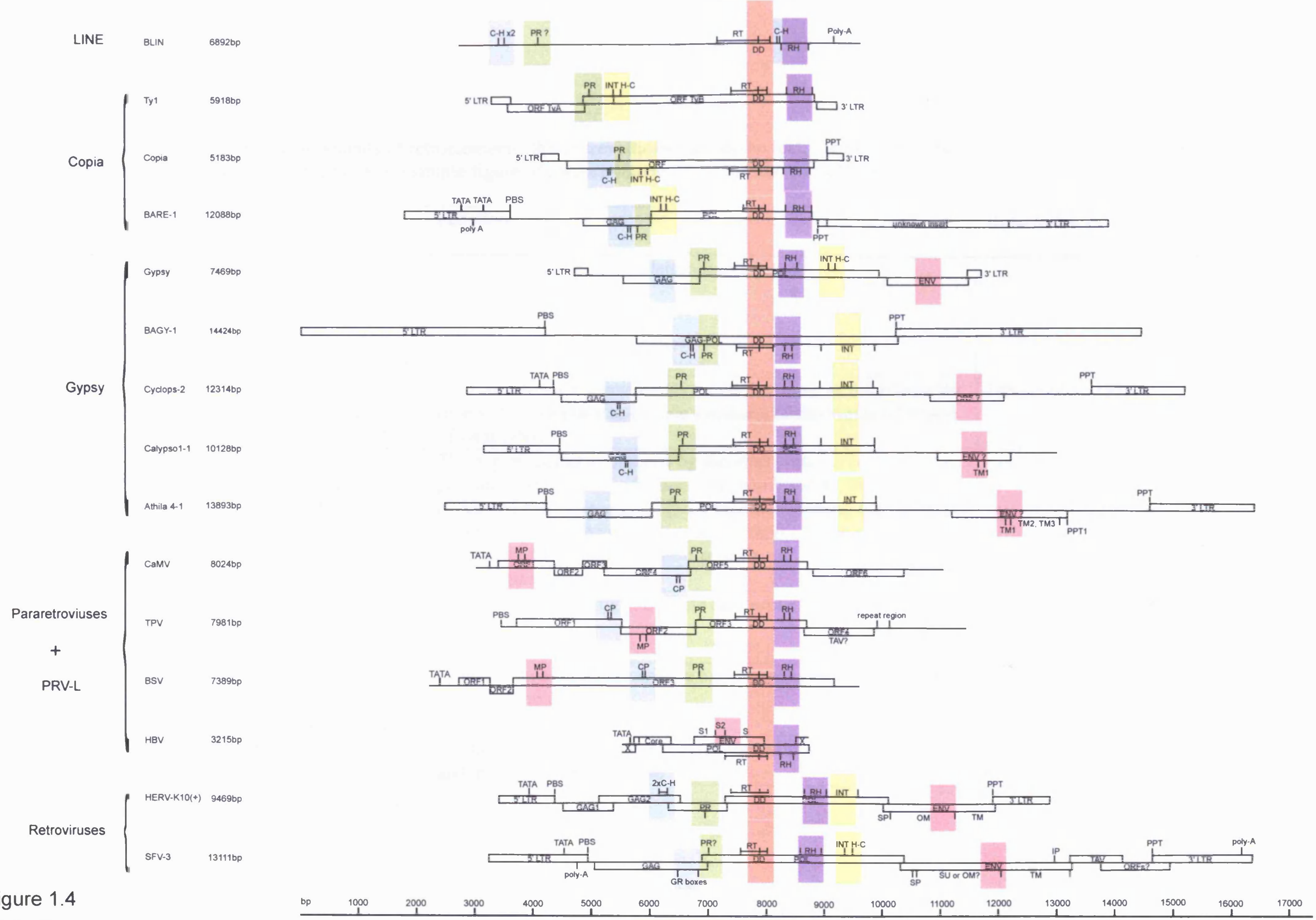
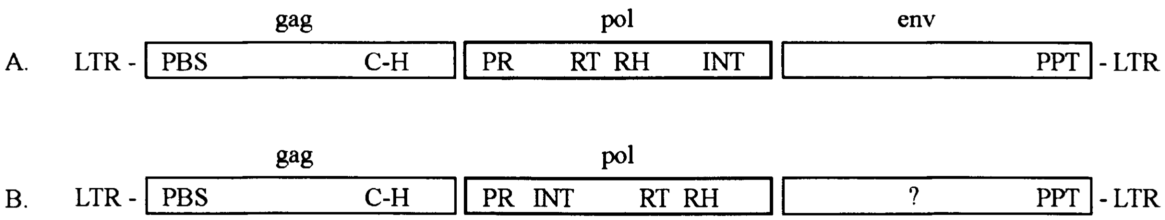

LINE
BLIN 6892bp
Copia
Ty1 5918bp
Copia 5183bp
BARE-1 12088bp
Gypsy
Gypsy 7469bp
BAGY-1 14424bp
Cyclops-2 12314bp
Calypso1-1 10128bp
Athila 4-1 13893bp
Pararetroviuses
+
PRV-L
CaMV 8024bp
TPV 7981bp
BSV 7389bp
HBV 3215bp
Retroviruses
HERV-K10(+) 9469bp
SFV-3 13111bp
C-H x2
PR ?
RT
C-H
DD
RH
Poly-A
5' LTR
3' LTR
ORF TyA
ORF TyB
PR
INT H-C
PPT
ORF
C-H
TATA
PBS
poly A
GAG
POL
unknown insert
ENV
GAG-POL
INT
ORF?
ENV ?
TM1
TM2, TM3
PPT1
MP
ORF1
ORF2
ORF3
ORF4
ORF5
ORF6
CP
repeat region
TAV?
S1
S2
S
Core
X
2xC-H
GAG1
GAG2
SP
OM
TM
PR?
GR boxes
poly-A
SU or OM?
IP
TAV
ORFs?
bp
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000

Figure 1.4

Table 1.1. Genes and other components of retroelements, the abbreviations used in the text, the full name, their position in the element and the function are listed. Below the table is a simple figure of a *gypsy*-like and a *copia*-like retroelement.

Gene or component	Full name	Position	Function	References
ORF	Open reading frame		Sequence capable of translation into a protein	
LTR	Long terminal repeat	Flankings of *copia*, *gypsy* and retrovirus	Contain the promotor	
PBS	Primer binding site	About 18 nt at the end of the 5'LTR	Binding site for a specific tRNA that functions as the primer for reverse transcriptase to initiate synthesis of the minus (-) strand of viral DNA	Petropoulos, 1997
gag	Group-specific antigen	Usually one of the first ORFs	The gag precurser is cleaved by the viral protease (encoded by pol) into three mature products: the matrix (MA), the capsid (CA), and the nucleocapsid (NC) together forming the "capsid" which surrounds the genome – this complex is the virus core. Equivalent to the coat or transit protein.	Lecellier and Saïb, 2000.
CP	Coat protein		Equivalent to gag	
Cys-His or C-H	Cysteine-histidine repeat motif	C-terminal of gag	RNA or DNA binding site of the coat protein or gag	de Kochko *et al.*, 1998
GR box		C-terminal of gag in certain retroelements	Contains three glysine/arginine basic sequences – functionally equivalent to Cys-His?	Lecellier and Saïb, 2000
pol	Polyprotein		Contains aspartic protease, reverse transcriptase and RNAse H and in some cases integrase	

PR	Aspartic protease	pol	Cleaves the full length mRNA. PR has a significant role in the processing of the polyprotein precursor into the mature form.	Ono *et al.*, 1986
RT	Reverse transcriptase	pol	RNA dependant DNA polymerase – translates RNA to DNA	
RH	Ribonuclease H/ RNase H	pol	RNase H is an enzyme that specifically degrades RNA hybridized to DNA.	Ohtani, 1999 (abs only)
INT	Integrase	pol	Enzyme responsible for removing two bases from the end of the LTR and inserting of the linear double stranded DNA copy of the retroelement genome into the host cell DNA	Petropoulos, 1997
env	Envelope gene	After pol, but not in pararetrovirus if MP=env	Envelope genes mediate the binding of virus particles to their cellular receptors enabling virus entry, the first step in a new replication cycle. Thus the envelope genes give retroelements the ability to spread between cells and individuals - infectivity. Contain the proteins SU (surface) and TM (transmembrane).	Löwer *et al.*, 1996
MP	Movement protein		Cell to cell movement, equivalent to env?	Hull, 2002
TAV	Transactivator		Regulating translation of the polycistronic mRNA	de Kochko *et al.*, 1998
PPT	Polypurine tract	7-18 nt just upstream of the 3'LTR	The ppt produce the RNA primer for the synthesis of the plus (+) strand of viral DNA	Petropoulos, 1997



Example of position of genes and components in a *gypsy*-like (A.) and a *copia*-like (B.) retroelement.

1.4.1 A short description of the retroelements in figure 1.4.

See also figure 1.1 and 1.2 for classification of elements.

LINE element:

BLIN; a LINE element from barley (*Hordeum vulgare*) with a genome of about 6892 bp. It is too degenerate to give border sites for the possible two ORFs. It has three Cys-His motifs, the two first are in reading frame one corresponding to ORF1, followed by a possible protease motif, the third is in reading frame two, corresponding to ORF2 - between RT and RNase H. A poly-A motif is present at the end of the element. It has a GC content of 62% and is represented with 40-50 copies per genome (Vershinin *et al.*, 2002; EMBL AJ270056; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

Copia elements:

Tyl; a *copia* element from yeast (*Saccharomyces cerevisiae*) with a genome of 5918 bp. There are 5' and 3' LTRs with 97% identity. It contains two ORFs, TyA and TyB. TyB contains protease, integrase, RT and RNase domains. The GC content is 37% (Boeke *et al.*, 1988; EMBL M18706; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

Copia; an element from *Drosophila melanogaster* with a genome size of 5183 bp. The element is bordered by LTRs having 97% identity. There is only one ORF. There is a Cys-His and a protease motif in the first third of the sequence followed by an integrase motif, the RT and RNase domains are in the last third. The GC content is 33% (Emori *et al.*, 1985; EMBL X02599; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

BARE-1; a *copia* element from barley (*Hordeum vulgare*) with a genome size of 12,088 bp. *BARE*-1 is bordered by LTRs with 96% identity. PBS is complementary to the 3' end of the wheat initiator methionyl-tRNA. PPT is present after *pol*. There are one or two ORFs followed by an insert of unknown origin and function. The *gag* has a Cys-His and a protease motif, *pol* has integrase, RT and RNase H. The whole element has a GC content of 47% (Manninen and Schulman, 1993; EMBL Z17327)

Gypsy elements:

Gypsy; an element from *Drosophila melanogaster* with a genome of 7469 bp. It is bordered by short LTRs, 49% identical. There are three ORFs, *gag*, *pol* and *envelope*. *Pol* contains protease, RT, integrase and RNase H. The GC content is 46% (Petropoulos, 1997; NCBI AF033821; Xiong and Eickbush, 1990).

BAGY-1; a *gypsy* element from barley (*Hordeum vulgare*) with a genome size of 14,424 bp. The bordering LTRs are 94% identical. The PBS next to the 5' LTR is complementary to the 3' end of a methionin initiator tRNA from wheat. A PPT is present just upstream of the 3' LTR. One *gag-pol* ORF was designated containing Cys-His motif, protease, RT, RNase H and integrase. The GC content is 46% (Panstruga *et al.*, 1998; EMBL Y14573; Xiong and Eickbush, 1990; Wright and Voytas, 2002).

Cyclops-2; a *gypsy*-like element from pea (*Pisum sativum*) being 12,314 bp long. It is bordered by LTRs that are 95% identical. The PBS is probably complementary to tRNA-glu from pea. The PPT is next to the 3'LTR. There are three ORFs, *gag*, *pol* and one of unknown function. *Gag* has the Cys-His motif, *pol* has protease, RT, RNase H and integrase. The unknown ORF has no homology with known *envelope* genes of other retroelements and is surrounded by non-coding regions. The element is GC rich, 42% and present with about 500 copies (Chavanne *et al.*, 1998; EMBL AJ000640; Wright and Voytas, 2002).

Calypsol-1; a *gypsy*-like element from soybean (*Glycine max*) with a genome of 10,128 bp. It only has the 5'LTR intact with a PBS next to it. There are three ORFs, *gag*, *pol* and a putative *envelope*. *Gag* has the Cys-His motif, *pol* has protease, RT, RNase H and integrase. A transmembrane domain is found in the putative *envelope* ORF. The GC content is 44% (Wright and Voytas, 2002; EMBL AF186182).

Athila4-1; a *gypsy* like element from *Arabidopsis thaliana*, 13,893 bp long. The LTRs are 94% identical with PBS and PPT next to them. There is no Cys-His motif in the *gag* region. *Pol* has protease, RT, RNase H and integrase. There is a putative *envelope* gene surrounded by non-coding regions. Three transmembrane domains are found in the *envelope*-like ORF including a second PPT. The GC content is 43% (Wright and Voytas, 2002; EMBL AC007209; Malik and Eickbush, 2001).

Pararetroviruses and pararetrovirus-like sequences (PRV-L):
CaMV (*Cauliflower mosaic virus*); a *Caulimovirus* with a genome of 8024 bp. There are six ORFs and an intergenic region. Hull (2002) gives two additional small ORFs, ORF7 before ORF1 and ORF8 within ORF4, they seem to have no significance. The movement protein is located in ORF1, the coat protein in ORF4, and protease, RT and RNase H are located in ORF5. As with other caulimoviruses and badnaviruses the numbering of the sequence begins at the putative 5' minus-strand priming site, conserved tRNA-met. The GC content is 40% (Franck *et al.*, 1980; EMBL J02048; de Kochko *et al.*, 1998; Harper and Hull, 1998; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

TPV (Tobacco pararetrovirus); a PRV-L sequence from tobacco (*Nicotiana tabacum*) 7981 bp long. It has four ORFs plus a repeat and intergenic region. ORF1 contains the coat protein, ORF2 contains the movement protein and ORF3 encodes protease, RT and RNase H. ORF4 has a transactivator (TAV). The GC content is 28% (Jakowitsch *et al.*, 1999; EMBL NTA238747; de Kochko *et al.*, 1998; Harper and Hull, 1998; Xiong and Eickbush, 1990).

BSV (*Banana streak virus*); a *Badnavirus* with a genome of 7389 bp. There are three ORFs of which the third is very large (5.5 kb) and contains all the genes: movement protein, coat protein, protease, RT and RNase H. The function of ORF1 and ORF2 is unknown. All badnaviruses have two different Cys-His motifs in the CP region. The GC content is 41% (Harper and Hull, 1998; EMBL AJ002234; de Kochko *et al.*, 1998; Hull, 2002; Malik and Eickbush, 2001).

HBV (*Hepatitis B virus*); a hepadnavirus with a genome of 3215 bp. In HBV nt 1 is set to be at an *Eco*RI restriction site. In the effort to adjust HBV to the other sequences it was decided to place a start point at the site for initiation of viral DNA synthesis at nt 1611 which are then nt 1. There are four ORFs. The core is equivalent to the coat protein of other pararetroviruses. The *envelope* ORF encodes three polypeptides possibly with transmembrane function. The *pol* contains RT and RNase

H. The function of ORF x is unknown but it is able to activate many viral and cellular promoters as well as several signal transduction pathways. The GC content is about 48%. Hepadnaviruses do not encode protease (Takahashi *et al.*, 1998; EMBL AB014360; Seeger, 1999; Hull, 1999; Xiong and Eickbush, 1990; Malik and Eickbush, 2001)

Retroviruses:

HERV-K10(+); a human endogenous retrovirus with a genome of 9469 bp. The LTRs are 99% identical with adjacent PBS and PPT. The PBS is complementary to tRNAlys. The element has five ORFs. Two *gag*, the second has two Cys-His motifs. The third ORF is designated protease. *Pol* has RT, RNase H and integrase. The *envelope* ORF has three transmembrane domains SP, OM and TM. HERV-K provirus (integrated form) is present with about 50 copies per haploid human genome. HERV-K10(+) is a prototype HERV-K genome as it is a construct of HERV-K10 plus a 290 bp fragment from HERV-K8 which is deleted from HERV-K10. Although defective in *gag* and *envelope* this virus still serves as a useful standard for sequence comparison. The GC content is 42% (Ono *et al.*, 1986; EMBL M14123; Manninen and Schulman, 1993; Löwer *et al.*, 1996; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

SFV-3 (*Simian foamy virus*); a spumavirus isolated from an African green monkey with a genome of 13,111 bp. The element is bordered by LTRs 100% identical with adjacent PBS and PPT. There are three large and three small ORFs. The classical tripartite retroviral division of *gag* does not exist in foamy viruses (see table 1.1 *gag*) and *gag* contain instead some GR-boxes (glysine/argenine) complementary to the Cys-His motif seen in other retroelements. *Pol* contains a potential protease, RT and integrase. The *envelope* gene has three transmembrane domains SP, SU/OM and TM including and internal promoter (IP) used as a second site of initiation of the plus (+) strand during reverse transcription. After the *envelope* is an ORF containing a putative TAV followed be two small ORFs of unknown function. The GC content is 38% (Renne *et al.*, 1992; EMBL M74895; Lecellier and Saïb, 2000; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

Many of the above mentioned elements have LTRs with a \ery high similarity which could mean that they have replicated relatively recently. Except for *Gypsy* having only 50% similarity between the 5' and 3' LTR. All of the elements except for the LINE have a GC content of less than 50%.

1.4.2 Conserved regions of retroelements

In addition to the general outline of the retroelements in figure 1.4 detailed structures of the most conserved domains such as the Cys-His motif, protease, RT, RNase H and integrase were compared. The conserved areas for each domain was found and the sequences aligned in Clustal W, additionally manual adjustment of the alignments are shown and conserved areas highlighted following finds of other researchers.

The RT domain contains the most overall conserved sequences often used for alignment and comparisons of groups of elements. In figure 1.5 sequences are aligned corresponding to domain 3-7 in Xiong and Eickbush (1990). The most conserved DD motif (D = aspartic acid) is used as fixed point; additional well-conserved motifs are marked, as is less well-conserved local motifs. BLIN and HBV have longer sequence spans than the others and a portion of amino acids have been removed and replaced with the corresponding number, 22 for BLIN and 53 for HBV, both at the start of the alignment.

The RNase H domain is used less often for alignment and sequence comparison (figure 1.6). RNase H is part of the polyprotein and is used to degrade RNA in RNA/DNA hybrids. Malik and Eickbush (2001) made an alignment of RNase H within retroelements and highlighted some single amino acids believed to be important in the catalytic reaction of the protein; D, E, D, D (E = glutamic acid). For the sampled elements this can be written as $DX_{27-48}EX_{18-33}DX_{29-54}D$ (where X = any amino acid) to show the distance between conserved amino acids. Additionally the non-LTR retrotransposons and retroviruses have an H (H = histidine) between the last two Ds. A DXS motif (S = serine) can be detected in many of the sequences including several other single or multiple amino acids, all or some of which are found in the retroelements in figure 1.4. McClure (1991) analysed a longer RNase H alignment.

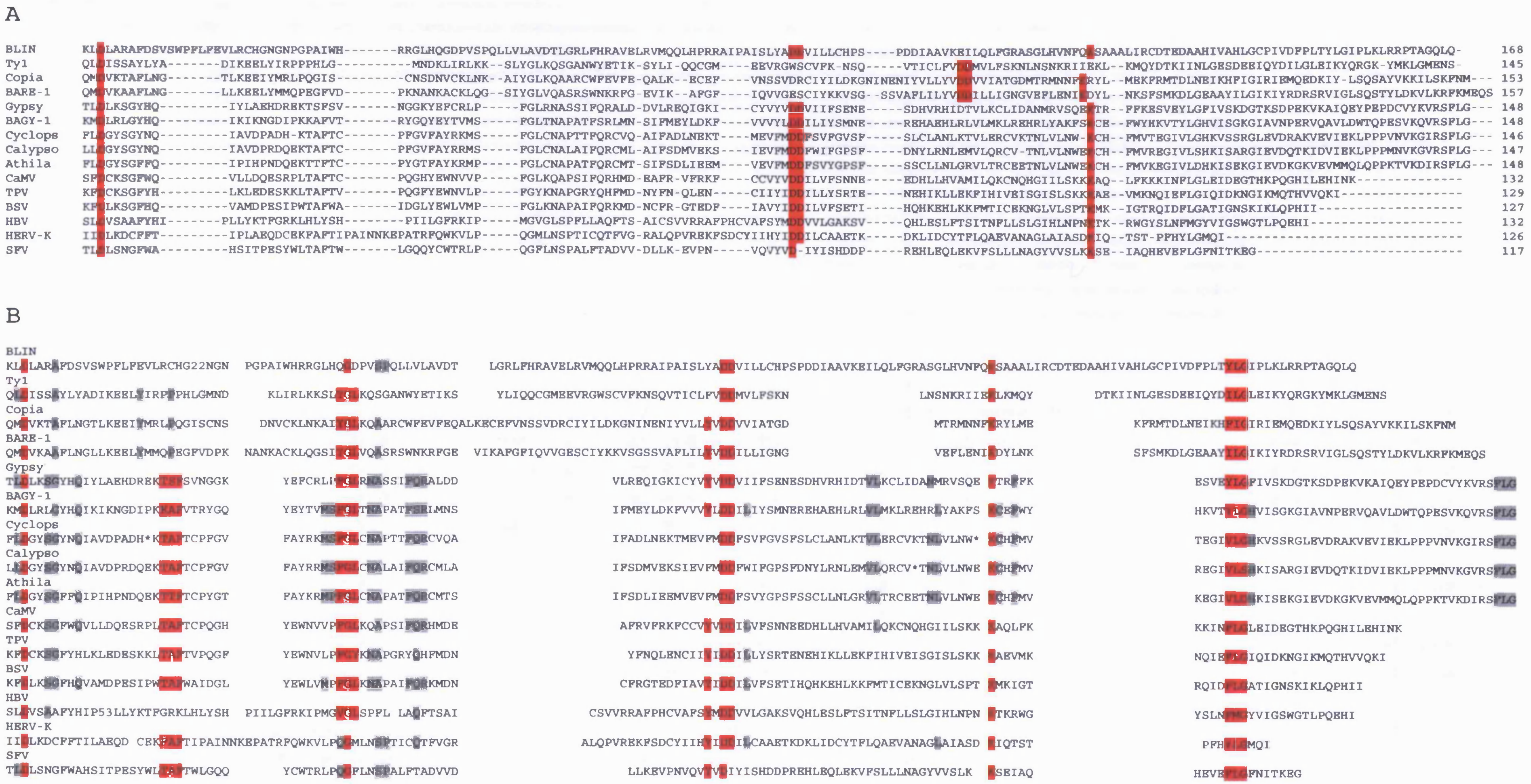


Figure 1.5. Alignment of the conserved reverse transcriptase region (RT) of all the retroelements in figure 1.4. The sequences cover what correspond to domain 3-7 in Xiong and Eickbush (1990). A. Clustal W alignment (Gap open, 5; End gap, 10; Gap ext, 0.05; gap distance, 4). B) manual alignemnt For references see sequence description.

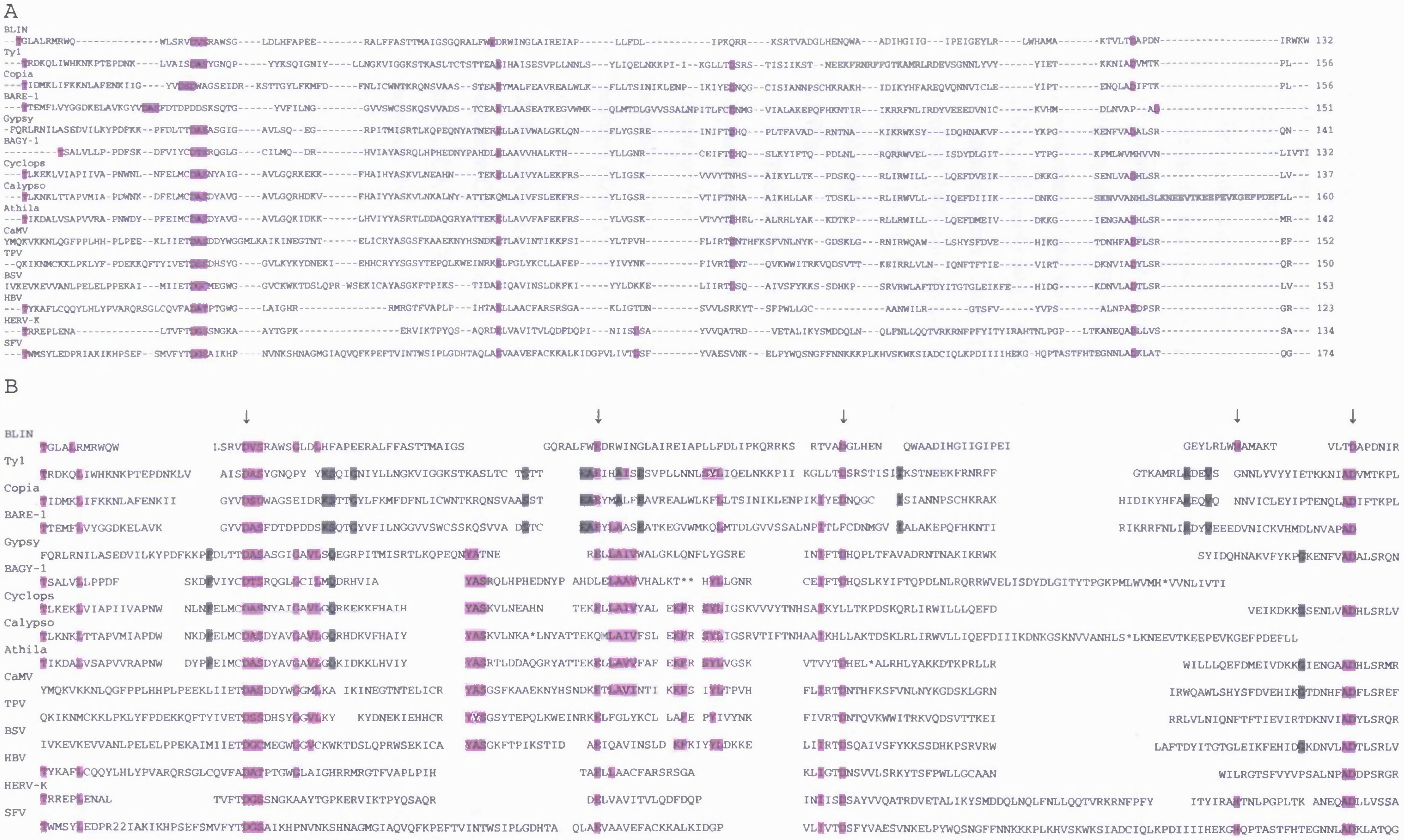
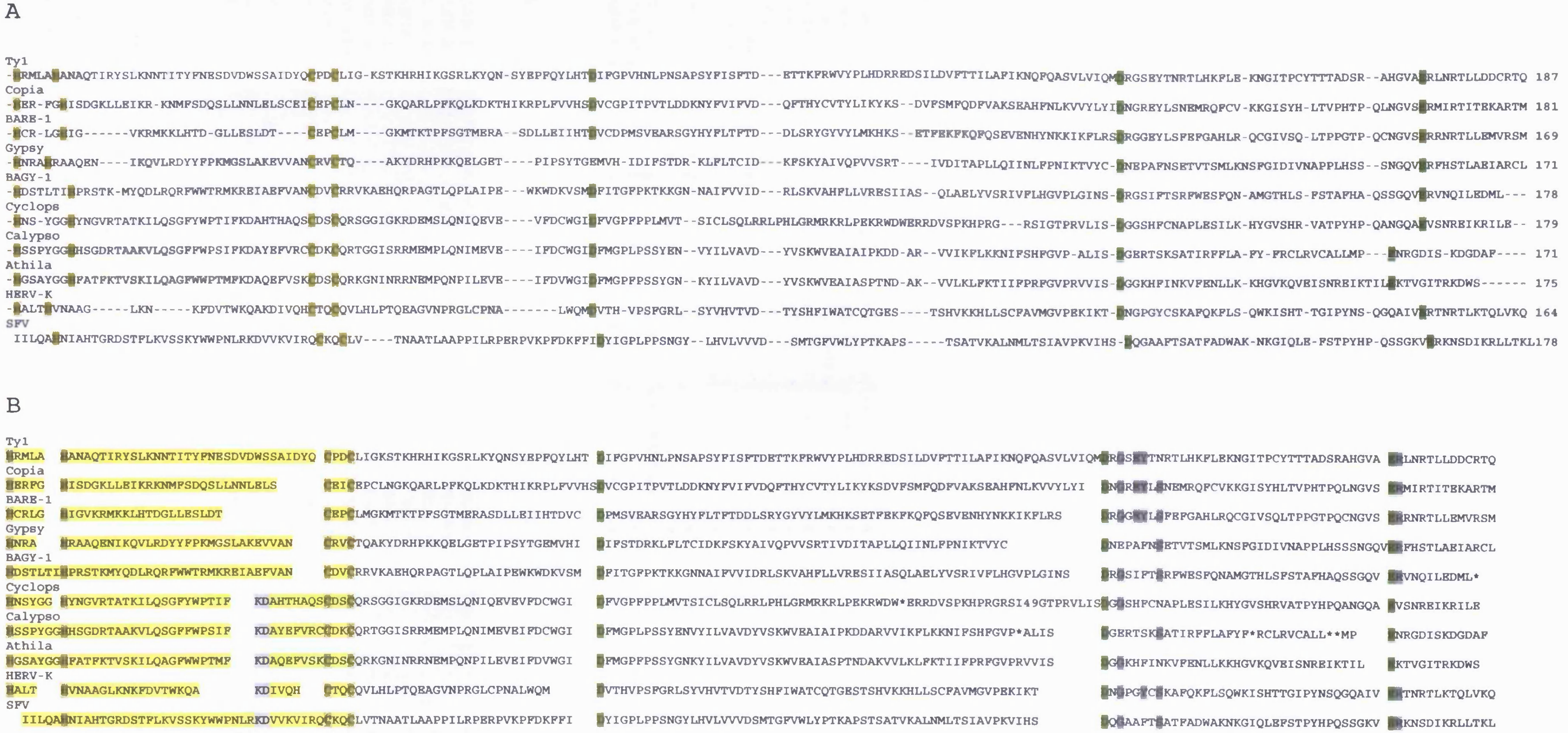


Figure 1.6. Alignment of the RNase H region (RH) of the retroelements in figure 1.4. A. Clustal W alignment (Gap open, 5; End gap, 10; Gap ext, 0.05; gap distance, 4). B) Manual alignment. The arrows above the alignment point to amino acid residues believed to be important for the catalytic mechanisms of RNase H; D, E, D, (H), D. For references see description of sequences.

The integrase is part of the polyprotein and mediates the integration of an element into the nuclear DNA of the host. The integrase domain contains both a well-conserved zinc finger (HHCC) (C = cystein) and a $DDX_{35}E$ motif, usually called DD35E (Khan *et al.*, 1991; Fayet *et al.*, 1990). These two motifs were found in the *copia*, *gypsy* and retrovirus elements in figure 1.4. In these elements the zinc finger motif became $HX_{3-6}HX_{20-33}CX_2C$. SFV is missing the first H. (figure 1.7). Additionally, the two amino acids KD (K = lysine) are conserved between *Cyclops*, *Calypso*, *Athila*, HERV-K and SFV. The DD35E motif is some 26-32 amino acids downstream of the last C in the zinc finger. In figure 1.7 the general motif is $DX_{52-64}DX_{32-36}E$, excluding *Cyclops* which has a very long sequence between the two Ds - 111 amino acids. Part of this sequence has been shortened and replaced with numbers (49) corresponding to the amount of amino acids removed to allow the alignment to fit. A few other less conserved amino acids are marked. Capy *et al.* (1996) found no similarities to the integrase domain of LTR retrotransposons in LINEs.

The Cystein-Histidine motif at the C-terminal of *gag* or in the coat protein is very well conserved (Covey, 1986) and found in LINE, *copia* and *gypsy* elements, in pararetroviruses and a retrovirus (figure 1.8). The function is not very clear but possible it binds genomic RNA or DNA to assist in packaging of virus particles and perhaps other processing. It consists of a short sequence with a characteristic pattern of cystein and histidine amino acids, making up a zinc finger. The motifs in the LINE, *copia*, *gypsy* and retrovirus elements are very similar with the amino acid sequence $CX_2CX_{3-4}HX_4C$ whereas the pararetroviruses have an additional CX in the motif, $CXCX_2CX_4HX_4C$. The third LINE Cys-His motif has longer intervals between the C and H than the two other LINE motifs $CX_4CX_5HX_6C$. The second Cys-His motif in BSV is rather different from the other having six Cs and an H, $CX_2CX_7HX_3CX_2CX_4CX_2C$.

Aspartic protease is part of the polyprotein and its function is to cleave the full length mRNA. The protease region is poorly conserved, the best homology being an $LX_{0-4}DXG$ motif (L = leucine and G = glycine), the rest being widely spaced single amino acids (figure 1.9). In some cases there were more than one sequence from an element having a suitable fit. See also McClure (1991).



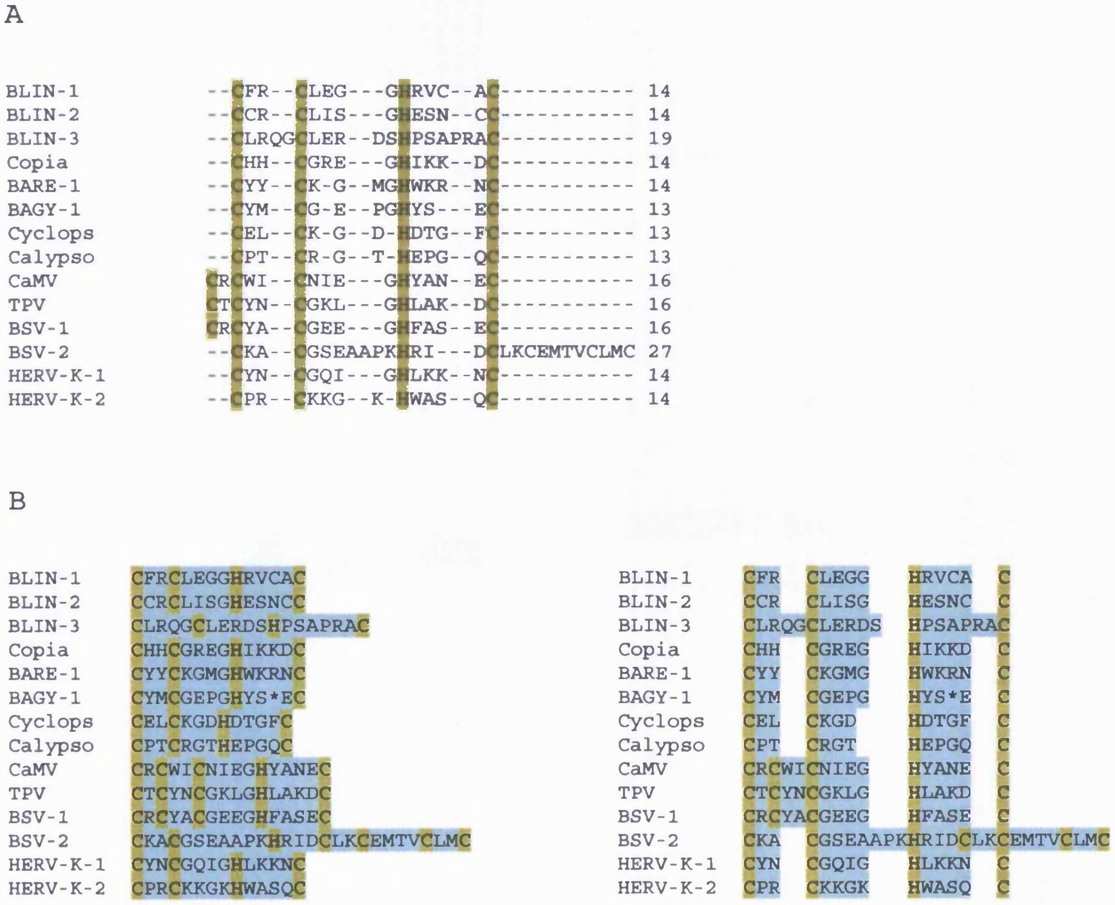
A

```
Ty1
-HRMLAHANAQTIRYSLKNNTITYFNESDVDWSSAIDYQCPDCLIG-KSTKHRHIKGSRLKYQN-SYEPFQYLHTDIFGPVHNLPNSAPSYFISFTD---ETTKFRWVYPLHDRREDSILDVFTTILAFIKNQFQASVLVIQMDRGSEYTNRTLHKFLE-KNGITPCYTTTADSR--AHGVAERLNRTLLDDCRTQ 187
Copia
-HER-FGHISDGKLLEIKR-KNMFSDQSLLNNLELSCEICEPCLN----GKQARLPFKQLKDKTHIKRPLFVVHSDVCGPITPVTLDDKNYFVIFVD---QFTHYCVTYLIKYKS--DVFSMFQDFVAKSEAHFNLKVVYLYIDNGREYLSNEMRQFCV-KKGISYH-LTVPHTP-QLNGVSERMIRTITEKARTM 181
BARE-1
-HCR-LGHIG------VKRMKKLHTD-GLLESLDT----CEPCLM----GKMTKTPFSGTMERA--SDLLEIIHTDVCDPMSVEARSGYHYFLTFTD---DLSRYGYVYLMKHKS--ETFEKFKQFQSEVENHYNKKIKFLRSDRGGEYLSFEFGAHLR-QCGIVSQ-LTPPGTP-QCNGVSERRNRTLLEMVRSM 169
Gypsy
-HNRAHRAAQEN----IKQVLRDYYFPKMGSLAKEVVANCRVCTQ----AKYDRHPKKQELGET----PIPSYTGEMVH-IDIFSTDR-KLFLTCID---KFSKYAIVQPVVSRT----IVDITAPLLQIINLFPNIKTVYC-DNEPAFNSETVTSMLKNSFGIDIVNAPPLHSS--SNGQVERFHSTLAEIARCL 171
BAGY-1
-HDSTLTIHPRSTK-MYQDLRQRFWWTRMKREIAEFVANCDVCRRVKAEHQRPAGTLQPLAIPE---WKWDKVSMDFITGFPKTKKGN-NAIFVVID---RLSKVAHFLLVRESIIAS--QLAELYVSRIVFLHGVPLGINS-DRGSIFTSRFWESFQN-AMGTHLS-FSTAFHA-QSSGQVERVNQILEDML--- 178
Cyclops
-HNS-YGGHYNGVRTATKILQSGFYWPTIFKDAHTHAQSCDSCQRSGGIGKRDEMSLQNIQEVE----VFDCWGIDFVGPFPPLMVT--SICLSQLRRLPHLGRMRKRLPEKRWDWERRDVSPKHPRG---RSIGTPRVLIS-DGGSHFCNAPLESILK-HYGVSHR-VATPYHP-QANGQAEVSNREIKRILE-- 179
Calypso
-HSSPYGGHHSGDRTAAKVLQSGFFWPSIFKDAYEFVRCCDKCQRTGGISRRMEMPLQNIMEVE----IFDCWGIDFMGPLPSSYEN--VYILVAVD---YVSKWVEAIAIPKDD-AR--VVIKFLKKNIFSHFGVP-ALIS-DGERTSKSATIRFFLA-FY-FRCLRVCALLMP---ENRGDIS-KDGDAF---- 171
Athila
-HGSAYGGHFATFKTVSKILQAGFWWPTMFKDAQEFVSKCDSCQRKGNINRRNEMPQNPILEVE----IFDVWGIDFMGPFPSSYGN--KYILVAVD---YVSKWVEAIASPTND-AK--VVLKLFKTIIFPRFGVPRVVIS-DGGKHFINKVFENLLK-KHGVKQVEISNREIKTILEKTVGITRKDWS------ 175
HERV-K
-HALTHVNAAG-----LKN-----KFDVTWKQAKDIVQHCTQCQVLHLPTQEAGVNPRGLCPNA-------LWQMDVTH-VPSFGRL--SYVHVTVD---TYSHFIWATCQTGES----TSHVKKHLLSCFAVMGVPEKIKT-DNGPGYCSKAFQKFLS-QWKISHT-TGIPYNS-QGQAIVERTNRTLKTQLVKQ 164
SFV
 IILQAHNIAHTGRDSTFLKVSSKYWWPNLRKDVVKVIRQCKQCLV----TNAATLAAPPILRPERPVKPFDKFFIDYIGPLPPSNGY--LHVLVVVD---SMTGFVWLYPTKAPS-----TSATVKALNMLTSIAVPKVIHS-DQGAAFTSATFADWAK-NKGIQLE-FSTPYHP-QSSGKVERKNSDIKRLLTKL178
```

B

```
Ty1
HRMLA   HANAQTIRYSLKNNTITYFNESDVDWSSAIDYQ  CPDCLIGKSTKHRHIKGSRLKYQNSYEPFQYLHT  DIFGPVHNLPNSAPSYFISFTDETTKFRWVYPLHDRREDSILDVFTTILAFIKNQFQASVLVIQMDRGSEYTNRTLHKFLEKNGITPCYTTTADSRAHGVA  ERLNRTLLDDCRTQ
Copia
HERFG   HISDGKLLEIKRKNMFSDQSLLNNLELS        CEICEPCLNGKQARLPFKQLKDKTHIKRPLFVVHSDVCGPITPVTLDDKNYFVIFVDQFTHYCVTYLIKYKSDVFSMFQDFVAKSEAHFNLKVVYLYI   DNGREYLSNEMRQFCVKKGISYHLTVPHTPQLNGVS  ERMIRTITEKARTM
BARE-1
HCRLG   HIGVKRMKKLHTDGLLESLDT               CEPCLMGKMTKTPFSGTMERASDLLEIIHTDVC   DPMSVEARSGYHYFLTFTDDLSRYGYVYLMKHKSETFEKFKQFQSEVENHYNKKIKFLRS           DRGGEYLSFEFGAHLRQCGIVSQLTPPGTPQCNGVS  ERRNRTLLEMVRSM
Gypsy
HNRA    HRAAQENIKQVLRDYYFPKMGSLAKEVVAN      CRVCTQAKYDRHPKKQELGETPIPSYTGEMVHI   DIFSTDRKLFLTCIDKFSKYAIVQPVVSRTIVDITAPLLQIINLFPNIKTVYC                  DNEPAFNSETVTSMLKNSFGIDIVNAPPLHSSSNGQVERFHSTLAEIARCL
BAGY-1
HDSTLTIHPRSTKMYQDLRQRFWWTRMKREIAEFVAN      CDVCRRVKAEHQRPAGTLQPLAIPEWKWDKVSM   DFITGFPKTKKGNNAIFVVIDRLSKVAHFLLVRESIIASQLAELYVSRIVFLHGVPLGINS          DRGSIFTSRFWESFQNAMGTHLSFSTAFHAQSSGQV  ERVNQILEDML*
Cyclops
HNSYGG  HYNGVRTATKILQSGFYWPTIF    KDAHTHAQSCDSCQRSGGIGKRDEMSLQNIQEVEVFDCWGI   DFVGPFPPLMVTSICLSQLRRLPHLGRMRKRLPEKRWDW*ERRDVSPKHPRGRSI49GTPRVLISDGGSHFCNAPLESILKHYGVSHRVATPYHPQANGQA  EVSNREIKRILE
Calypso
HSSPYGGHHSGDRTAAKVLQSGFFWPSIF    KDAYEFVRCCDKCQRTGGISRRMEMPLQNIMEVEIFDCWGI   DFMGPLPSSYENVYILVAVDYVSKWVEAIAIPKDDARVVIKFLKKNIFSHFGVP*ALIS            DGERTSKSATIRFFLAFYF*RCLRVCALL**MP     ENRGDISKDGDAF
Athila
HGSAYGGHFATFKTVSKILQAGFWWPTMF    KDAQEFVSKCDSCQRKGNINRRNEMPQNPILEVEIFDVWGI   DFMGPFPSSYGNKYILVAVDYVSKWVEAIASPTNDAKVVLKLFKTIIFPRFGVPRVVIS            DGGKHFINKVFENLLKKHGVKQVEISNREIKTIL    EKTVGITRKDWS
HERV-K
HALT    HVNAAGLKNKFDVTWKQA        KDIVQH    CTQCQVLHLPTQEAGVNPRGLCPNALWQM       DVTHVPSFGRLSYVHVTVDTYSHFIWATCQTGESTSHVKKHLLSCFAVMGVPEKIKT              DNGPGYCSKAFQKFLSQWKISHTTGIPYNSQGQAIV  ERTNRTLKTQLVKQ
SFV
  IILQAHNIAHTGRDSTFLKVSSKYWWPNLRKDVVKVIRQCKQCLVTNAATLAAPPILRPERPVKPFDKFFI   DYIGPLPPSNGYLHVLVVVDSMTGFVWLYPTKAPSTSATVKALNMLTSIAVPKVIHS              DQGAAFTSATFADWAKNKGIQLEFSTPYHPQSSGKV  ERKNSDIKRLLTKL
```

Figure 1.7. Alignment of conserved regions from the integrase region (INT) of ten sequences from figure 1.4. A. Clustal W alignment (Gap open, 5; End gap, 10; Gap ext, 0.05; gap distance, 4). B. Manual alignment. The first motif (HHCC) is a zinc finger, from the present sequences it can be written as $HX_{3\text{-}6}HX_{20\text{-}33}CX_2C$. The next motif is DDE which, form the present sequences, can be written as $DX_{52\text{-}64}DX_{32\text{-}36}E$. For references see description of individual sequences.



A

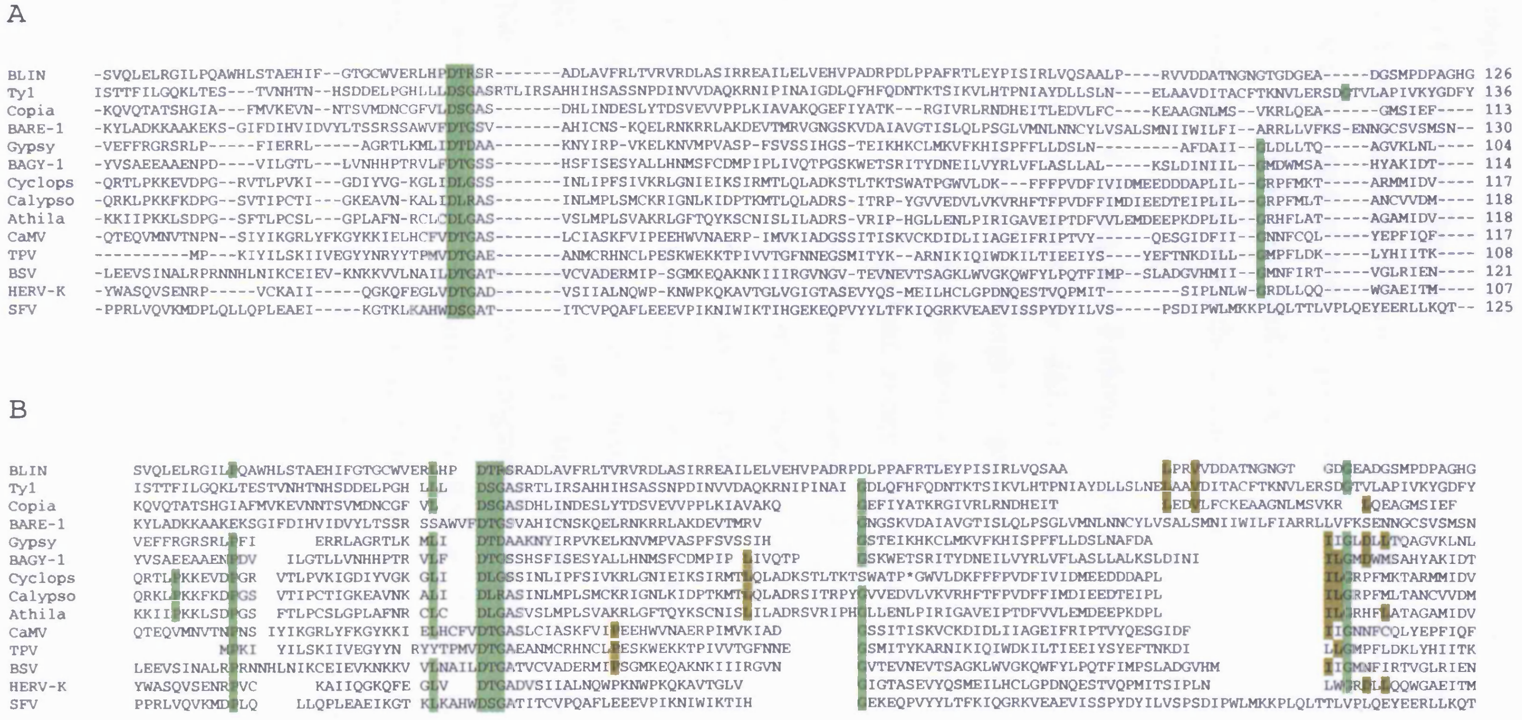
```
BLIN-1          --CFR--CLEG---GHRVC--AC----------- 14
BLIN-2          --CCR--CLIS---GHESN--CC----------- 14
BLIN-3          --CLRQGCLER--DSHPSAPRAC----------- 19
Copia           --CHH--CGRE---GHIKK--DC----------- 14
BARE-1          --CYY--CK-G--MGHWKR--NC----------- 14
BAGY-1          --CYM--CG-E--PGHYS---EC----------- 13
Cyclops         --CEL--CK-G--D-HDTG--FC----------- 13
Calypso         --CPT--CR-G--T-HEPG--QC----------- 13
CaMV            CRCWI--CNIE---GHYAN--EC----------- 16
TPV             CTCYN--CGKL---GHLAK--DC----------- 16
BSV-1           CRCYA--CGEE---GHFAS--EC----------- 16
BSV-2           --CKA--CGSEAAPKHRI---DCLKCEMTVCLMC 27
HERV-K-1        --CYN--CGQI---GHLKK--NC----------- 14
HERV-K-2        --CPR--CKKG--K-HWAS--QC----------- 14
```

B

```
BLIN-1      CFRCLEGGHRVCAC
BLIN-2      CCRCLISGHESNCC
BLIN-3      CLRQGCLERDSHPSAPRAC
Copia       CHHCGREGHIKKDC
BARE-1      CYYCKGMGHWKRNC
BAGY-1      CYMCGEPGHYS*EC
Cyclops     CELCKGDHDTGFC
Calypso     CPTCRGTHEPGQC
CaMV        CRCWICNIEGHYANEC
TPV         CTCYNCGKLGHLAKDC
BSV-1       CRCYACGEEGHFASEC
BSV-2       CKACGSEAAPKHRIDCLKCEMTVCLMC
HERV-K-1    CYNCGQIGHLKKNC
HERV-K-2    CPRCKKGKHWASQC
```

```
BLIN-1      CFR  CLEGG   HRVCA  C
BLIN-2      CCR  CLISG   HESNC  C
BLIN-3      CLRQGCLERDS  HPSAPRAC
Copia       CHH  CGREG   HIKKD  C
BARE-1      CYY  CKGMG   HWKRN  C
BAGY-1      CYM  CGEPG   HYS*E  C
Cyclops     CEL  CKGD    HDTGF  C
Calypso     CPT  CRGT    HEPGQ  C
CaMV        CRCWICNIEG   HYANE  C
TPV         CTCYNCGKLG   HLAKD  C
BSV-1       CRCYACGEEG   HFASE  C
BSV-2       CKA  CGSEAAPKHRIDCLKCEMTVCLMC
HERV-K-1    CYN  CGQIG   HLKKN  C
HERV-K-2    CPR  CKKGK   HWASQ  C
```

Figure 1.8. Cysteine-Histidine motif alignment with ten sequences from figure 1.4. A. Clustal W alignment (Gap open, 5; End gap, 10; Gap ext, 0.05; gap distance, 4). B) Mannual alignment, to the left shown as they are and to the right aligned. For references see description of sequences.



A

```
BLIN      -SVQLELRGILPQAWHLSTAEHIF--GTGCWVERLHPDTRSR-------ADLAVFRLTVRVRDLASIRREAILELVEHVPADRPDLPPAFRTLEYPISIRLVQSAALP----RVVDDATNGNGTGDGEA-----DGSMPDPAGHG 126
Ty1       ISTTFILGQKLTES---TVNHTN--HSDDELPGHLLLDSGASRTLIRSAHHIHSASSNPDINVVDAQKRNIPINAIGDLQFHFQDNTKTSIKVLHTPNIAYDLLSLN----ELAAVDITACFTKNVLERSDGTVLAPIVKYGDFY 136
Copia     -KQVQTATSHGIA---FMVKEVN--NTSVMDNCGFVLDSGAS-------DHLINDESLYTDSVEVVPPLKIAVAKQGEFIYATK---RGIVRLRNDHEITLEDVLFC----KEAAGNLMS--VKRLQEA------GMSIEF---- 113
BARE-1    -KYLADKKAAKEKS-GIFDIHVIDVYLTSSRSSAWVFDTGSV-------AHICNS-KQELRNKRRLAKDEVTMRVGNGSKVDAIAVGTISLQLPSGLVMNLNNCYLVSALSMNIIWILFI--ARRLLVFKS-ENNGCSVSMSN-- 130
Gypsy     -VEFFRGRSRLP-----FIERRL-----AGRTLKMLIDTDAA-------KNYIRP-VKELKNVMPVASP-FSVSSIHGS-TEIKHKCLMKVFKHISPFFLLDSLN---------AFDAII--GLDLLTQ-----AGVKLNL---- 104
BAGY-1    -YVSAEEAAENPD----VILGTL---LVNHHPTRVLFDTGSS-------HSFISESYALLHNMSFCDMPIPLIVQTPGSKWETSRITYDNEILVYRLVFLASLLAL-----KSLDINIIL--GMDWMSA-----HYAKIDT---- 114
Cyclops   -QRTLPKKEVDPG--RVTLPVKI---GDIYVG-KGLIDLGSS-------INLIPFSIVKRLGNIEIKSIRMTLQLADKSTLTKTSWATPGWVLDK---FFFPVDFIVIDMEEDDDAPLIL--GRPFMKT-----ARMMIDV---- 117
Calypso   -QRKLPKKFKDPG--SVTIPCTI---GKEAVN-KALIDLRAS-------INLMPLSMCKRIGNLKIDPTKMTLQLADRS-ITRP-YGVVEDVLVKVRHFTFPVDFFIMDIEEDTEIPLIL--GRPFMLT-----ANCVVDM---- 118
Athila    -KKIIPKKLSDPG--SFTLPCSL---GPLAFN-RCLCDLGAS-------VSLMPLSVAKRLGFTQYKSCNISLILADRS-VRIP-HGLLENLPIRIGAVEIPTDFVVLEMDEEPKDPLIL--GRHFLAT-----AGAMIDV---- 118
CaMV      -QTEQVMNVTNPN--SIYIKGRLYFKGYKKIELHCFVDTGAS-------LCIASKFVIPEEHWVNAERP-IMVKIADGSSITISKVCKDIDLIIAGEIFRIPTVY------QESGIDFII--GNNFCQL-----YEPFIQF---- 117
TPV       ----------MP---KIYILSKIIVEGYYNRYYTPMVDTGAE-------ANMCRHNCLPESKWEKKTPIVVTGFNNEGSMITYK--ARNIKIQIWDKILTIEEIYS----YEFTNKDILL--GMPFLDK-----LYHIITK---- 108
BSV       -LEEVSINALRPRNNHLNIKCEIEV-KNKKVVLNAILDTGAT-------VCVADERMIP-SGMKEQAKNKIIIRGVNGV-TEVNEVTSAGKLWVGKQWFYLPQTFIMP--SLADGVHMII--GMNFIRT-----VGLRIEN---- 121
HERV-K    -YWASQVSENRP-----VCKAII-----QGKQFEGLVDTGAD-------VSIIALNQWP-KNWPKQKAVTGLVGIGTASEVYQS-MEILHCLGPDNQESTVQPMIT--------SIPLNLW-GRDLLQQ-----WGAEITM---- 107
SFV       -PPRLVQVKMDPLQLLQPLEAEI-----KGTKLKAHWDSGAT-------ITCVPQAFLEEEVPIKNIWIKTIHGEKEQPVYYLTFKIQGRKVEAEVISSPYDYILVS-----PSDIPWLMKKPLQLTTLVPLQEYEERLLKQT-- 125
```

B

```
BLIN         SVQLELRGILPQAWHLSTAEHIFGTGCWVERLHP  DTRSRADLAVFRLTVRVRDLASIRREAILELVEHVPADRPDLPPAFRTLEYPISIRLVQSAA          LPRVVDDATNGNGT    GDGEADGSMPDPAGHG
Ty1          ISTTFILGQKLTESTVNHTNHSDDELPGH LLL   DSGASRTLIRSAHHIHSASSNPDINVVDAQKRNIPINAI GDLQFHFQDNTKTSIKVLHTPNIAYDLLSLNELAAVDITACFTKNVLERSDGTVLAPIVKYGDFY
Copia        KQVQTATSHGIAFMVKEVNNTSVMDNCGF VL    DSGASDHLINDESLYTDSVEVVPPLKIAVAKQ        GEFIYATKRGIVRLRNDHEIT           LEDVLFCKEAAGNLMSVKR   LQEAGMSIEF
BARE-1       KYLADKKAAKEKSGIFDIHVIDVYLTSSR SSAWVFDTGSVAHICNSKQELRNKRRLAKDEVTMRV          GNGSKVDAIAVGTISLQLPSGLVMNLNNCYLVSALSMNIIWILFIARRLLVFKSENNGCSVSMSN
Gypsy        VEFFRGRSRLPFI      ERRLAGRTLK MLI   DTDAAKNYIRPVKELKNVMPVASPFSVSSIH         GSTEIKHKCLMKVFKHISPFFLLDSLNAFDA                   IIGLDLLTQAGVKLNL
BAGY-1       YVSAEEAAENPDV   ILGTLLVNHHPTR VLF   DTGSSHSFISESYALLHNMSFCDMPIP LIVQTP      GSKWETSRITYDNEILVYRLVFLASLLALKSLDINI              ILGMDWMSAHYAKIDT
Cyclops      QRTLPKKEVDPGR  VTLPVKIGDIYVGK GLI   DLGSSINLIPFSIVKRLGNIEIKSIRMTLQLADKSTLTKTSWATP*GWVLDKFFFPVDFIVIDMEEDDDAPL                  ILGRPFMKTARMMIDV
Calypso      QRKLPKKFKDPGS  VTIPCTIGKEAVNK ALI   DLRASINLMPLSMCKRIGNLKIDPTKMTLQLADRSITRPYGVVEDVLVKVRHFTFPVDFFIMDIEEDTEIPL                  ILGRPFMLTANCVVDM
Athila       KKIIPKKLSDPGS  FTLPCSLGPLAFNR CLC   DLGASVSLMPLSVAKRLGFTQYKSCNISLILADRSVRIPHGLLENLPIRIGAVEIPTDFVVLEMDEEPKDPL                  ILGRHFLATAGAMIDV
CaMV         QTEQVMNVTNPNS IYIKGRLYFKGYKKI ELHCFVDTGASLCIASKFVIPEEHWVNAERPIMVKIAD        GSSITISKVCKDIDLIIAGEIFRIPTVYQESGIDF               IIGNNFCQLYEPFIQF
TPV                   MPKI  YILSKIIVEGYYN RYYTPMVDTGAEANMCRHNCLPESKWEKKTPIVVTGFNNE       GSMITYKARNIKIQIWDKILTIEEIYSYEFTNKDI               LLGMPFLDKLYHIITK
BSV          LEEVSINALRPRNNHLNIKCEIEVKNKKV VLNAILDTGATVCVADERMIPSGMKEQAKNKIIIRGVN        GVTEVNEVTSAGKLWVGKQWFYLPQTFIMPSLADGVHM            IIGMNFIRTVGLRIEN
HERV-K       YWASQVSENRPVC      KAIIQGKQFE GLV   DTGADVSIIALNQWPKNWPKQKAVTGLV            GIGTASEVYQSMEILHCLGPDNQESTVQPMITSIPLN             LWGRDLLQQWGAEITM
SFV          PPRLVQVKMDPLQ    LLQPLEAEIKGT KLKAHWDSGATITCVPQAFLEEEVPIKNIWIKTIH           GEKEQPVYYLTFKIQGRKVEAEVISSPYDYILVSPSDIPWLMKKPLQLTTLVPLQEYEERLLKQT
```

Figure 1.9. Alignment of the aspartic protease region (PR) of the retroelements in figure 1.4. A. Clustal W alignment (Gap open, 5; End gap, 10; Gap ext, 0.05; gap distance, 4). B) manual alignment. For references see description of sequences.

1.4.3 Relationship, origin and classification of retroelements

Summarizing the previous paragraphs it has been shown that retroelements probably have an ancient origin and that there are several regions which are conserved across retroelements, particularly the RT. The structure of retroelements indicates that various components or functions have been exchanged between element types during the course of evolution. This lack of strict borders between element families makes classification and evolution difficult to elucidate. Two solutions have been given in figure 1.1 and 1.2 and a third is given in figure 1.10 which provides natural (phylogenetic, functional or evolutionary) relationships between the different element classes. In the following the differences and weakness of these figures will be discussed.

The model of Xiong and Eickbush (1990) is more than ten years old and many independent discoveries have added to the knowledge about the function and relationship of retroelements although the general picture is the same. Phylogenetic analysis using the RNase H domain instead of the RT domain (Malik and Eickbush, 2001) gave the same overall pattern, except that the retroviruses are sister to LINE-like elements. Analysis of the RNase H domain showed that all elements on the left hand branch of figure 1.2 (except retroviruses) lacked an amino acid residue present in LINE elements. Therefore it was proposed that an early lineage of retroviruses replaced their existing RNase H domain with one from a LINE-like element. The structure of retroviruses supports this theory (see figure 1.4), as there is a gap between RT and RNase H which is larger than for the other elements. Remnants of the old RNase H site are thought to strengthen the RT-RNase H activity (Malik and Eickbush, 2001). Sequence analysis of the RNase H domain also confirm that caulimoviruses are much more similar to the *gypsy* retroelements than the hepadnaviruses (Malik and Eickbush, 2001). However, possibly RT and RNase H regions are too closely related to allow confirmation of the similarities. But not many other domains have or can be compared over all retroelements. Some has been done with the integrase domain (Capy *et al.*, 1996 and 1997) but basically only covers *copia*, *gypsy* and retroviruses. These authors suggest that the integrase domain with the DDE motif of LTR-retrotransposons and retroviruses originated from the transposases of some DNA transposable elements (indicated on figure 1.10).

An evolution scheme developed from figure 1.2 can be seen in figure 1.10, showing possible development and addition of components for early retroelements.

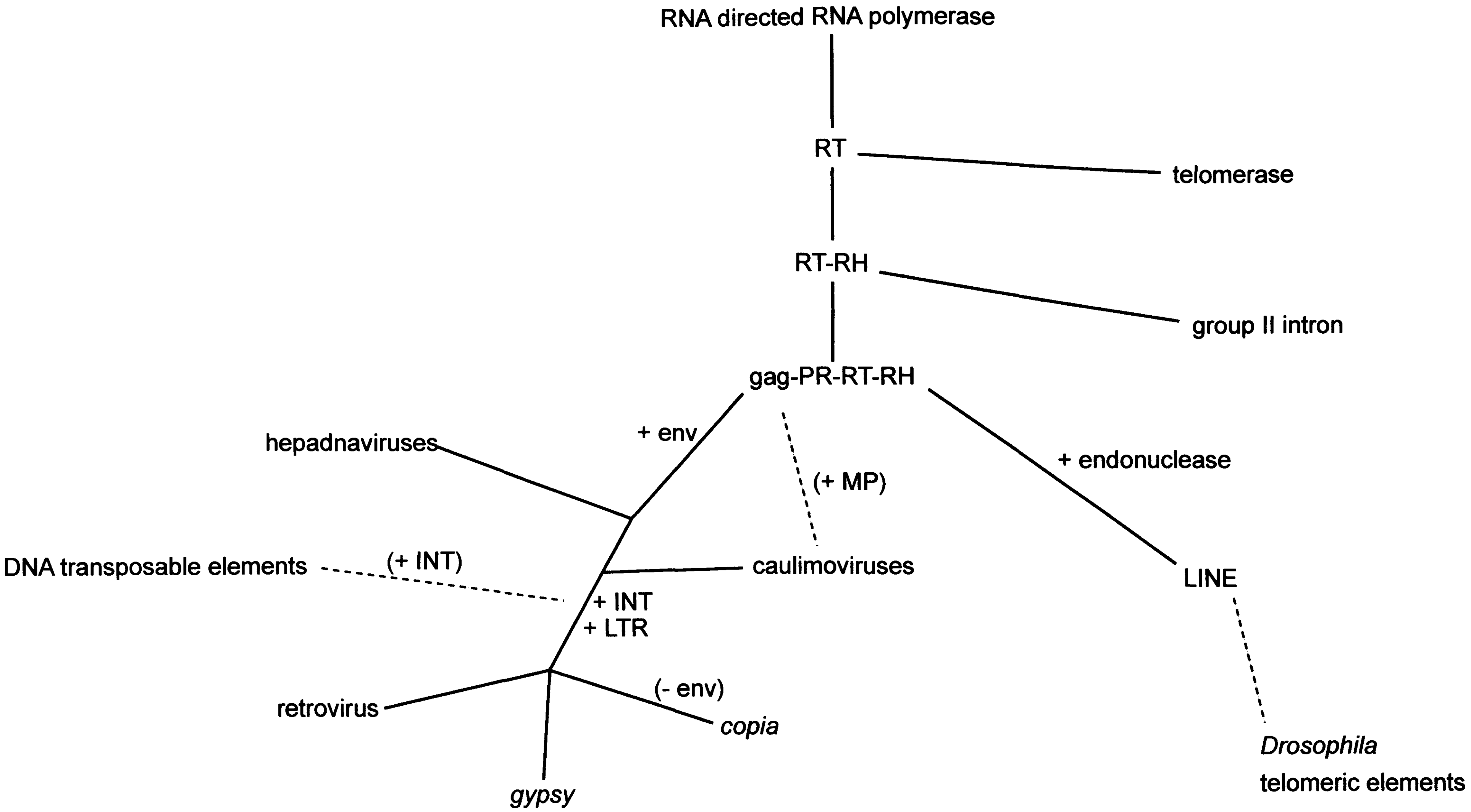


Figure 1.10. Possible evolution of retroelements and telomerase based on maximum parsimony of components.

The telomerase is shown as a branch from RT although it has been debated whether telomerase evolved from non-LTR retrotransposons or vice versa (Malik *et al.*, 1999). Placing group II introns on the branch after LINEs (figure 1.2) implies that bacteria acquired the RT after the eukaryote- prokaryote split. Recent speculation suggests that the group II introns might be older than LINEs (Eickbush, 1999; Malik and Eickbush, 2001). Therefore the branch to group II introns may be placed in a more ancient position (figure 1.10) developing from a RT-RNase H composition. The LINEs are given a separate branch from where the *Drosophila* telomeric elements possible developed (as suggested by Capy *et al.*, 1997). LTR-retrotransposons and retroviruses are probably closely related as they have a similar structure and replication cycle. The evolution of hepadnaviruses and caulimoviruses are not very clear and their positions on the branch are unsure. Caulimoviruses could branch off somewhere on the branch leading to *gypsy* or they may have developed from a more ancient element with the addition of MP, or perhaps caulimoviruses were developed from *gypsy*-like elements. Figure 1.10 is given a basis of maximum parsimony in domain composition and components but as mentioned above a more flexible evolution of retroelements with exchange of component are just as likely.

The integrated PRV-L sequences described in this study, and other similar viral sequences would, following figure 1.1, be placed in the viral group, and although originally infective they may have lost the ability. Placing them on figure 1.2 and 1.10 should be somewhere in extension of or branching from caulimoviruses. Considering caulimoviruses are relatively closely related to gypsy retrotransposons for the RT (figure 1.2) would it be surprising to find them widely integrated in plant genomes? But why and how should they loose the integrase? Maybe as Xiong and Eickbush (1990) suggests, by a virus capturing part of an LTR element.

Figure 1.1 indicates that elements with very similar structure and behaviour (retroviruses and *gypsy* elements) are present on two different branches along with less related sequences and that LINEs are closely related to the LTR-retrotransposons. In contrast figure 1.2 and 1.10 suggests LINEs and LTR-retrotransposons diverged early on from each other and that LTR retrotransposons are on the same branch as retroviruses. In support of the latter figure 1.4 shows that the structure of *gypsy* elements and retroviruses are very similar. The two families of pararetroviruses placed in the same suborder in figure 1.1 are wide apart in figure

1.2, the hepadnaviruses branching off early and the caulimoviruses being more closely related to the gypsy elements.

Given the evident flexibility in assembly of components into a 'retroelement', and the widespread distribution of different elements among living organisms essentially over the whole of evolutionary time, jumping of components between element types, homogenisation of sequences, and interspecies transfer of new elements are all real possibilities which will confound any attempt at making a coherent, monophyletic phylogeny. It is clear that some retroelement component parts - RT, for example - are remarkably conserved during evolution, but others, such as the *envelope* genes show minimal recognizable similarity in sequence, so they may be polyphyletic in origin.

1.5 Occurrence and activation of endogenous viral retro-elements

Evolution of copy number in retroelements can be interpreted as showing periods of low and high amplification and insertion activity, with evidence that this is related to the development or environment of the host (Grandbastien, 1998; Kalendar *et al.*, 2000). The retroviruses with LTRs and integrase domain have integration into nuclear DNA as a part of their replication cycle whereas the pararetroviruses replicate outside of the nuclear DNA with no nuclear integration (Hull and Covey, 1996). In recent years several pararetroviruses have nevertheless been found in nuclear DNA (Harper *et al.*, 2002). Unusually high activity or unexpected integration is often found in connection with stress events such as tissue culture and wide hybridizations (Dahal *et al.*, 2000; Lockhart *et al.*, 2000; Mhiri *et al.*, 1997). Endogenous (host genome integrated) retroviruses have long been known in human (International Human Genome Consortium, 2001).

1.5.1 Endogenous retroviruses

Endogenous retroviral sequences (ERV) have now been identified in more than 25 vertebrate orders across six classes including sharks, bony fish, amphibians, reptiles, birds, primates, rodents, carnivore, rabbit-like animals (Lagomorpha), "hoofed animals" (Perissodactyla) and "cloven-footed animals" (Artiodactyla) (Benit *et al.*, 1999; Herniou *et al.*, 1998). Some of the retroviral groups may be restricted to particular vertebrate classes and many of the sequences contained stop codons and frame shifts which may indicate that these sequences are old and not active (Herniou *et al.*, 1998). In human there are several different families of human endogenous retroviruses but most of them are highly defective (Löwer *et al.*, 1996). The biologically most active endogenous retrovirus family HERV-K has retained long ORFs and the capacity to be expressed at the RNA and protein level inducing an immune response, but no exogenous virus has been detected (Löwer *et al.*, 1996). HERV-K can be traced back to at least the divergence of old and new world monkey, 25 million years ago.

Waugh O'Neill *et al.* (1998) report on the interspecific cross between two wallabies which resulted in an offspring with a large amount of an endogenous retrovirus KERV-1 together with undermethylation and chromosome remodelling of the genome. The retrovirus was found confined to the centromeres by *in situ* hybridization but not detectable in either parent.

In plants, no sequences resembling retroviruses where the infective form is RNA have yet been reported. However, new evidence suggests the PRV-L sequences, with a DNA integrating molecule, may be present in plants.

1.5.2 Endogenous pararetroviruses

Hepadnaviridae is the family of mammalian pararetroviruses of which human *Hepatitis B virus* (HBV) is the best-known example. HBV infects liver tissue and infection is linked to cancer. Several cases have shown that HBV can become illegitimately integrated in the host genome (Pineau *et al.*, 1996; Tagieva *et al.*, 1995; Wang *et al.*, 2001) and that there might be a preferred splice site in the virus genome. Integration is not an essential step in the replication cycle and may be correlated with carcinoma (cancer cells).

1.5.3 Endogenous plant pararetrovirus-like sequences

Genome integrated PRV-L sequences have been reported from *Musa*, *Petunia* hybrids, *Nicotiana*, *Lycopersicon* and *Oryza* between 1996 and 2002 and some evidence from epidemiology and molecular biology suggests that there is a possibility that the sequences can be expressed and give rise to episomal viruses and infection. This mode of integration contrast with the normal replication cycle of pararetroviruses.

Bananas (*Musa*) can be infected by BSV (Dahal *et al.*, 1998) and is causing disease worldwide especially in subsistence bananas and plantains. However, in some *Musa* cultivars the symptoms do not appear to correspond with the presence of an insect vector in the field, and infection seems to be severe following times of stress. Especially progeny from tissue culture of certain varieties have a very high incidence of the disease (Dahal *et al.*, 2000). Evidence from PCR amplification, *in situ* hybridization (Harper *et al.*, 1999) and genomic library screening (Ndowora *et al.*, 1999) indicated that there was a BSV-like sequence in the nuclear DNA of some *Musa* varieties. DNA fibre-stretch fluorescent *in situ* hybridization showed that the integrated BSV sequence is repeated in two different structures of 150 kb and 50 kb respectively (Harper *et al.*, 1999). It is believed that sexual hybridization and tissue culture can give rise to episomal viruses by recombination of the integrated sequence. The integrated viral sequence shares 99.6% identity with the sequenced episomal virus (Ndowora *et al.*, 1999). *Musa* consists of A and B genome diploids species and a number of auto- and allopolyploid varieties. Geering *et al.* (2001) found that there is more than one species of BSV integrated in the genome of *Musa* and remnants of other BSV sequences are found in both A and B genomes. These data about integration of BSV sequences into the *Musa* genome has lead to suggestions that consideration and care is needed for not only the safe movement of germplasm but also in breeding and tissue culture (Harper and Hull, 1998).

Integrated PVCV sequences have been detected in *Petunia hybrida* (*P. axillaris* ssp. *axillaris* x *P. integrifolia* ssp. *inflata*), possibly in high copy number (Richert-Pöggeler *et al.*, 1996). Although *in situ* hybridization indicates that the sequences are concentrated at relatively few sites (Richert-Pöggeler, Schwarzacher and Harper unpublished data). There is evidence that a complete PVCV genome is present in one *Petunia* cultivar and that several smaller sequences are present in many cultivars (Harper *et al.*, 2002). The presence of the integrated virus sequence is correlated with the appearance of disease symptoms and virus particles in some *P. hybrida* varieties, again under particular environmental conditions.

The allohexaploid *Nicotiana edwardsonii* (Lockhart *et al.*, 2000) was formed by the hybridization between *N. clevelandii* (female, 4x) and *N. glutinosa* (male, 2x). In *N. edwardsonii*, the spontaneous presence of episomal virus particles (TVCV) was discovered under certain environmental conditions. Positive Southern hybridization of TVCV to genomic DNA of *N. edwardsonii* and *N. glutinosa* suggested that TVCV was integrated in the nuclear DNA. It was concluded that the occurrence of episomal TVCV was possibly triggered by the interspecific hybridization and the succeeding chromosome doubling. *In situ* hybridization indicated that the copy number of the integrated sequence was in the order of thousands, and at a large number of loci (Lockhart and Schwarzacher, unpublished data). Only one of the parents, *N. glutinosa*, was found to have integrated sequences resembling the virus, although virus particles were not detected.

In some cases, sequences with clear homology to PRV-L sequences but with no evidence for viral expression have been reported. In a study of DNA flanking transgenes, Jakowitsch *et al.* (1999) found several sequences with high homology to known pararetroviruses. From these fragments it was possible to assemble a hypothetical 7981 bp PRV-L genome called TPV. TPV is most closely related to TVCV (75%) and CsVMV (42%) at the nucleotide level and has the same genomic structure as TVCV. No episomal virus was detected. TPV is further described in chapter III.

Genomic sequence data also suggest the presence of PRV-L sequences in other species. Budiman *et al.* (2000), with the long-term aim of sequencing the tomato genome, was generating a sequence-tagged connector (STC) framework from a BAC library of *Lycopersicon esculentum*. Retrotransposon sequences were common among the data but there were also some PRV-L sequences. Fragments of RTBV, especially the RT-RNase H region are found in the rice genome, and a smaller fragment homologous to a part of the movement protein is also found. For instance EMBL AP000559 (Sasaki *et al.*, 1999 unpublished data). Additionally a survey on a STC framework from rice detected three PRV-L fragments with relationship to RTBV (Mao *et al.*, 2000).

1.5.4 Other endogenous plant viruses

Another case of an integrated plant virus has been reported although it is not retroelement related. The integrant is most related to *Tomato golden mosaic virus* (*Begomovirus*, *Geminiviridae*) which consists of two single stranded circular DNA molecules. The integrated fraction called geminivirus related DNA (GRD) is a hairpin loop (replication recognition) and a defective replicase (Bejarano *et al.*, 1996).

In the tobacco (*Nicotiana tabacum*) genome GRD sequences were located by in situ hybridization on the long arm of a pair of small submetacentric chromosomes from the T-genome (Kenton *et al.*, 1995). The insert was recognised as distinct sets of approximately 25 multiple direct repeats. Results were confirmed with Southern and slot blots giving an estimated copy number of 360 GRD repeats in the tobacco genome (Bejarano *et al.*, 1996). Ashby *et al.* (1997) found multiple copies of the GRD sequences in the genomes of related *Nicotiana* species; *N. tomentosiformis*, *N. tomentosa* and *N. kawakamii*. It does not occur in nine more distantly related *Nicotiana* species nor in various other solanaceous and non-solanaceous plants.

Phylogenetic analysis strongly suggested that the GRD sequences were descended from an ancient new-world geminivirus - as *N. tabacum* too is a new-world plant. The organization of GRD sequences resembles the pattern of duplications and rearrangements that occur by illegitimate recombination of artificially introduced DNA within eukaryotic chromosomes. No activity or expression of GRD was detected. Bejarano *et al.* (1996) consider if plants which have a high plasticity and propagate easily from cuttings are more likely to have incorporated exogenous DNA.

From the cases mentioned above it is seen that many forms of viruses are integrated in the nuclear genomes of a wide selection of eukaryotes. In several cases they are found to be active either by transcription or by emergence of episomal virus particles. It is possible that the presence of virus-like sequences might have an influence on the host genome. Bejarano *et al.* (1996) speculate that possible reasons for the persistence of the GRD sequence in the genome of some *Nicotiana* species could be active selection e.g. co-segregation because of integration proximal to a favoured allele. Alternatively the integration could alter the expression of neighbouring host genes in a useful way or by giving geminivirus resistance by anti-sense expression. Similar speculations could be addressed to the other integrated viruses.

1.6 Interaction with the host

The interaction between retroelement and host genome is difficult to interpret. Whether or not endogenous retroelements are beneficial to the host is questionable; it is more likely that organisms that have been able to exploit or effectively control insertion and activity of retroelements may have had a higher survival rate. Insertion of retroelements can cause changes in the host genome such as insertional mutation, chromosome breakage, chromosome rearrangement, altered gene regulation and sequence amplification. Even remnants of integrated viruses can still have an effect on the host. They can have promoter/enhancer activity in the LTRs, active splice sites, ORFs or RT activity and have the ability to be retrotransposed (Löwer, 1999).

1.6.1 Cellular reverse transcriptase

Retroelements are not the only genomic component with RT function. The enzyme telomerase, which adds DNA segments at chromosome ends in most eukaryotes, was found to have motifs in common with retroelement RT including the DD motif and some other conserved single amino acids (Lingner *et al.*, 1997). Further data show that the eukaryotic telomerase contains its own RNA template as an integral part of the enzyme, and a given eukaryotic species has a characteristic telomeric DNA sequence. The RNA template works by binding to an existing single stranded chromosome end and synthesizing a short DNA sequence by reverse transcription. The template disassociates and moves one step forward to repeat the synthesis (Blackburn 1992 and 2000). The relationship of telomerase to retroelements have been debated (Eickbush, 1997) among other because of the unusual situation seen in *Drosophila* where chromosome ends are not extended by telomerase but by repeated insertion of two LINE-like retrotransposons (Pardue *et al.*, 1997).

Another relationship between a retroelement and a "normal" cellular component are the group II introns. Dai and Zimmerly (2002) speculate that group II introns (introns removed from RNA by a selfsplicing mechanism) might be the origin of spliceosomal introns (removed by a large RNA-protein spliceosome complex), which successfully invaded and colonized eukaryotic genomes.

1.6.2 Activity and host control

The original insertion and periods of activity of retroelements are difficult to estimate as some sort of standard is needed to estimate mutation rates and retroelements may behave differently from the host DNA. Thorough investigation of a portion of the maize genome and the inserted retrotransposons gave an estimate for the last transposition events (SanMiguel *et al.*, 1998). It was shown that all of the retrotransposons had inserted within the last six million years and most of them in the last three million years experiencing a burst of activity. The time of insertion was calculated by the divergence of the two LTRs as they are assumed to be identical at the time of insertion (SanMiguel *et al., 1998*). It is thought that a low copy number of these sequences had existed for a long time and for some reason they amplified quite recently as mentioned above (Bennetzen, 2000). Both *Athila* and *Tat1 gypsy* retrotransposons have high sequence degeneracy in the coding regions whereas each element has near sequence identity of their 5' and 3' LTRs (>95%). This also accounts for other retroelements, see paragraph 1.4.1. The high similarity of LTRs suggests that these elements integrated relatively recently or that transcripts from defective elements were acted upon in trans to generate the insertions (Wright and Voytas, 1998).

LTRs and internal domains of *BARE*-1 are conserved in the genus *Hordeum* and it was therefore possible to PCR amplify fragments and use these as probes on genomic DNA (Vicient *et al.*, 1999). It was found that the genome size was positively correlated with the copy number of *BARE*-1 internal domain. On average it was found that the ratio of internal *BARE*-1 region to LTR was 1 to 15 within the genus *Hordeum*. Combining these data it was seen that the more LTRs relative to internal *BARE*-1 element, the smaller the contribution of *BARE*-1 to the genome. Vicient *et al.* (1999) therefore suggests that recombination of LTRs and excision of the *BARE*-1 element is the plants way of limiting the number of retroelements.

Most interspersed repeats in mammals predate the eutherian (placental) radiation (70 million years ago). This shows the extremely slow rate with which non-functional sequences are cleared from vertebrate genomes. In the draft human sequence only three full-length copies of retroviruses can be identified with all ORFs intact (International Human Genome sequencing Consortium, 2001). The proportion of old elements is much larger in humans than in the *Drosophila*, *Caenorhabditis* or *Arabidopsis*. Perhaps genome sweeping mechanisms have removed most elements in these organisms with small genomes.

Transposon activity in the mouse genome has not undergone the decline seen in humans and proceeds at a much higher rate (International Human Genome sequencing Consortium, 2001). The mouse seems to have very active chromosomes/genome with many rearrangements (Gregory *et al.*, 2002) which may promote activity of integrated elements although it could be argued that it is the activity of integrated elements promoting chromosome rearrangements.

ERVs have predominantly been found to be expressed in germcells, embryonic tissue and the placenta from where they can be transferred to the offspring. A tobacco endogenous PRV-L transcript was also found to be active in shoot apical meristems of *Arabidopsis* (Löwer, 1999; Mette *et al.*, 2002).

1.6.3 Horizontal transfer

Most transposable elements are passed on from parent to offspring in a vertical manner. If a sequence is passed on to a non-offspring individual, another species, genus or family it is called horizontal transfer.

Horizontal transfer is often suggested when very high similarity between elements from related or distant species is observed or there are inconsistencies between the phylogeny of the element and that of the hosts (Capy *et al.*, 1994). For example the phylogenetic branching pattern of *gypsy*-like retroelements agrees roughly with the accepted overall phylogeny of their host organisms. However, *Cyclops* from *Pisum* is exceptional in that it does not cluster with plant retroelements but falls into the wider animal clade close to micropia, an LTR-retrotransposon from *Drosophila* (Chavanne *et al.*, 1998). The study of Friesen *et al.* (2001) also finds that *Cyclops* is different from the other plant *gypsy* elements and although this may indicate horizontal transmission, they speculate that it is actually a representative of a different sub-family of elements.

There might be other explanations to the observed phenomena than horizontal transfer. Transposable elements from a common ancestor could evolve differently (slow/rapid) in different genetic backgrounds and/or environments leading to the species seen today. If the common ancestor contained polymorphic elements they could be distributed differentially giving the result that diverse species have more similar elements than closer related species (Capy *et al.*, 1994).

The *P* element, a DNA transposable element, was believed to have been horizontally transferred from one species of *Drosophila*, *D. willistoni* to another species, *D. melanogaster*. The evidence was that the *P* element has a very patchy distribution in *Drosophila*, but it is lacking in old laboratory strains of *D. melanogaster* and it is practically identical in the two *Drosophila* species. The horizontal transfer of the *P* element may have been mediated by a virus (Daniels *et al.*, 1990).

The relative lack of observed horizontal transfers of elements between mammals (International Human Genome sequencing Consortium, 2001) was believed to be due to the well developed immune system, as horizontal transfer requires infectious vectors such as viruses, against which the immune system guards.

1.6.4 Silencing and resistance

Organisms have developed methods to keep the amount of transposable elements and viruses at an acceptable level for their own survival. These methods involve among other methylation of endogenous sequences and silencing of endo- and exogenous sequences by means of RNA sequences.

Methylation can lead to gene specific methylation or transcriptional gene silencing which possible prevents transcription. In many plant systems small pieces of RNA guide *de novo* methylation of homologous DNA sequences. It was proposed that the primary function of cytosine methylation in vertebrates might be defence against transposons and retroviruses as the large majority (some say >90%) of methylations are within transposons (Sharp, 2001). Methylation is most effectively targeted against promoters of transposable elements. Apart from this short-term protection, methylation also provides a potential mechanism for long-term protection by driving a C to T mutation of the element sequence (Bestor, 1999).

In post transcriptional gene silencing (PTGS) the gene is transcribed but no mRNA is accumulated, as it is quickly degraded. RNA interference (RNAi) is PTGS mediated by a homologous dsRNA sequence. The enzyme Dicer digests the dsRNA sequence is into small fragments of RNA (21-25 bp) called small interfering RNAs (siRNA) which together with Dicer becomes a nuclease complex. This complex can base pairs with homologous RNA sequences and degrade0 them so no protein product can be synthesized. A siRNA can have a double role by acting both in PTGS in the cytoplasm and in methylation in the nucleus. The dsRNA could possible derive from the infecting virus itself or a genome integrated sequence (Matzke *et al.*, 2001; Waterhouse *et al.*, 2001).

Mette *et al.* (2002) investigated whether TPV-L sequences exhibit features that would be compatible with a potentially new type of homology dependant virus resistance. Transgenes driven by the TPV-L transcriptional regulatory sequences were silenced and methylated when introduced into *Nicotiana tabacum*, but remained active and unmethylated in non-host species such as *Arabidopsis*, which are devoid of sequences homologous to TPV-L sequences. TPV-L sequences might supply resistance to a related virus - though this virus has not yet been detected. It was believed that stably methylated TPV-L sequences have supplied long-term viral immunity, perhaps accompanied by weakening or extinction of the related exogenous virus. According to Löwer (1999) free-living pathogenic particles analogous to endogenous sequences are found less often than expected. This is possibly because over evolutionary time, such particles have become incorporated in the host genome conferring resistance to the free-living form, which then gradually became extinct. But what about TVCV, closely related to TPV-L sequences, which does show up as an episomal virus in *N. tabacum*? Perhaps the plants of Mette *et al.* (2002) (and Jakowitsch *et al.*, 1999) were not exposed to the environmental conditions giving rise to an episomal virus. Or TVCV is too different from TPV to be regarded as its exogenous form.

Lately it has been suggested that the system of natural antiviral defence in the cells could be used to fight serious viral diseases such as HIV (Carmichael, 2002; Jacque *et al.*, 2002). Apparently the system works equally well with externally supplied siRNAs as well as siRNA produced by the cell itself. Until now silencing of HIV has only been tried on cell cultures where the major obstacles seems to be that an siRNA of 21-25 bp has to be 100% identical to the target to have a strong effect and that the viruses mutate quickly. So it seems like PTGS in some instances only works under very high homology between cellular sequence and virus. This stringency is probably necessary to avoid destruction of beneficial RNA.

1.7 Viruses as useful tools?

Viruses of all types have co-evolved with plant and animal genomes since the earliest period of the evolution of eukaryotes. As with other pathogens, most infections are benign and there is a balance between the virus and host cell, although some chronic infections may reduce productivity with relatively small symptoms and others kill their host. But the infectivity and adaptation of viruses can be used by humans to beneficial effect.

Plant viruses including caulimoviruses, geminiviruses and RNA viruses have been used as gene vectors to introduce and express foreign genes into plants. The advantages are that the recombinant virus can be introduced directly into a grown plant as an alternative to plant transformation and that the virus moves throughout the plant. The disadvantages are that the sequence of interest is non heritable so that a continuous inoculation is necessary in annuals. The virus can revert to wild type and give disease problems and the size of insert is rather limited (Hull, 2002). For this purpose vertebrate retroviruses are more efficient transfer vectors as the sequence gets integrated into the target genome. They have been used for gene therapy and marking of transplanted tissue (Miller, 1997), although recombination with existing endogenous retroviruses can be an undesired and potential fatal side effect. And a question is if immunosuppressive medication induces re-expression of ERVs in the recipient and/or in the transplanted tissue especially in the case of tissue from e.g. pig or monkey to human (Löwer, 1999).

Several plant virus promoters have been tested and used for expression in non-host plants often with the purpose of driving transgene inserts. From CaMV (Pauk *et al.*, 1995), BSV (Schenk *et al.*, 2001), CsVMV (Verdaguer *et al.*, 1998) and ScBV (sugarcane) (Tzafrir *et al.*, 1998). Also the untranslated leader sequence has been used as enhancers of mRNA translation (Hull, 2002).

Plant virus particles modified to expose certain peptides on the surface have been tried as vaccines. It should be possible to produce them in large amount at low cost (for developing countries), they can be given orally as food, they can be stored easily in e.g. fruits and they are inactive in animals. One of the problems is that the virus loses the insert (Hull, 2002). Gough *et al.* (1999) isolated peptides that bind specifically to *Cucumber mosaic virus* (CMV) coat protein, they can be displayed on carrier proteins and there are potential for using them for plant resistance.

Experiments have been made to use viruses in functional genomics - virus induced gene silencing. A gene incorporated in a virus vector will silence a homologous gene in plants (Hull, 2002).

Reverse transcriptases from retroviruses are routinely used in laboratories in RT-PCR.

It was found that people with HIV infection who were co-infected with the harmless Hepatitis G Virus (HBG) had a higher survival rate. Results suggested that HGV might impair HIV replication without causing any disease itself. The amount of

HGV increased in patients having antiretroviral therapy. Possible HGV infection is a marker for the presence of other factors that slow down the progression of HIV (Tillmann *et al.*, 2001; Xiang *et al.*, 2001)

Overall, the abilities of viruses, ranging from their control of cellular machinery to components such as coat proteins and RT, have provided many important tools for biotechnology. It is likely that study of new groups of viruses will lead to further tools and products for exploitation of plant genomes.

1.8 Plants studied

After this introduction follow chapters based on results from the study of PRV-L sequences in plants. Most work of this study has been based on potato, *Solanum tuberosum* 'Desiree', from the family Solanaceae. Potato was chosen as the basic model plant for the following reasons. The majority of reports on integrated PRV-L sequences have been to other genera of the family, TPV and TVCV from *Nicotiana* and PVCV from *Petunia*. Potato is the fourth most important food crop (FAO Statistical Database, http://apps.fao.org) and widely grown. Last but not least potato was positive in initial trials of selected primers.

The cultivated potato has a long history of growth and many wild relatives. The allotetraploid *Solanum tuberosum* was basically generated by crossing the first domesticated diploid potato with a wild diploid species. Continued selection and incorporation of desired traits has lead to the modern potato cultivars (Hawkes, 1990). Solanaceae is divided into several tribes. Potato and tomato (*Lycopersicon*) belong to the same tribe, Solaneae. *Nicotiana* and *Petunia* belong to Cestreae.

This study included plants from major groups in the plant kingdom and figure 1.11 shows a simplified cladogram of the relationship between these plants (see also table 2.1). Marchantiopsida (liverworts) is often recognized as the earliest class of land plants (Qiu *et al.*, 1998; Tree Of Life Project, tolweb.org/tree/phylogeny.html; Jacobsen and Jensen, 1997) closely followed by Bryopsida (mosses). In these two classes the gametophyte generation is what we see as the green plant while the following plants have the sporophyte generation as the dominating feature and are also referred to as tracheophytes. Mosses are then the sister group to tracheophytes (Judd *et al.*, 1999). Lycopodiopsida (club mosses), Equisetopsida (horsetails) and Polypodiopsida (ferns) reproduce by dispersing spores. The further developed seed plants can be divided into what is called Gymnosperms including Ginkgo, Pinidae (conifers) and Gnetidae and the Angiosperms. The Angiosperms are probably the most studied group of plants divided into Monocotyledons and Dicotyledons. Nymphales with *Nuphar* is recognized as one of the earliest Angiosperms (Qui *et al.*, 1999), not clearly falling into either Monocotyledon or Dicotyledon groups.

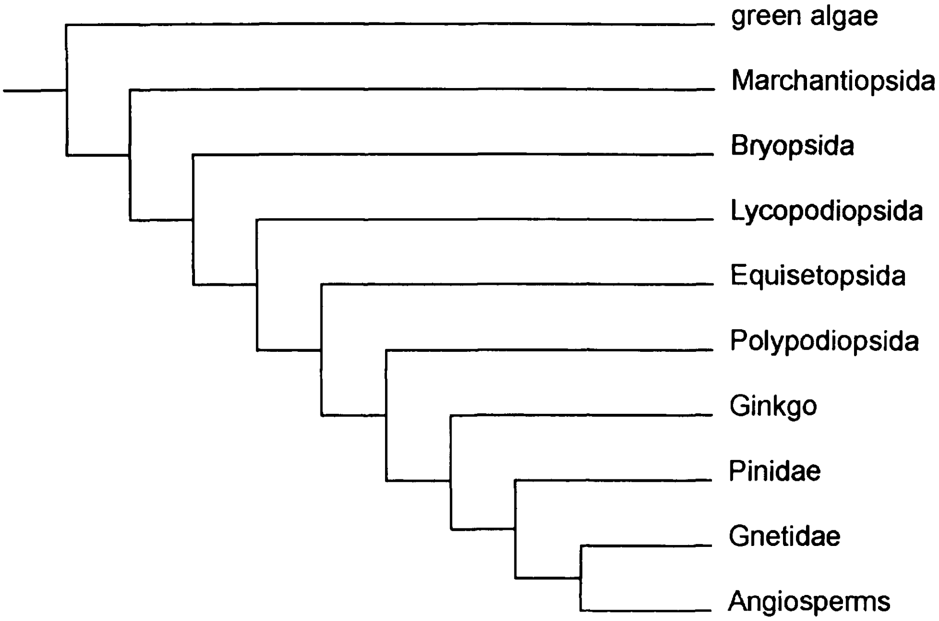


Figure 1.11. Simplified phylogeny showing the relationship of some major groups of land plants. Green algae has been placed as the sister group.

The DNA C-value is a good way of comparing genome sizes of plants and the amount of data is growing (http://www.rbgkew.org.uk/cval/homepage.html). The 1C-value corresponds to the DNA amount of pollen tetrads or the unreplicated haploid genome. DNA amount in plants and other eukaryotes are usually expressed in picogram (pg) or in megabase pairs (Mb): 1 pg=965Mb. Today flow cytometry is used to estimate C-value. The DNA 1C-value is very variable in Angiosperms differing by more than 600-fold, ranging from less than 0.2 pg in *Arabidopsis thaliana* to 127.4pg in *Fritillaria assyriaca*. These differences in genome size are largely due to varying amount of repetitive sequences (Bennett and Leitch, 1995). The tetraploid potato has 48 chromosomes (2n=4x=48) and the DNA content is rather small with a 1C value of 1.8 pg.

1.9 Aims

The review above gives the background to the evolution, phylogeny and components of plant pararetroviruses and retroelements in the context of viruses in p lants and animals and the evolution of genomes. In general terms, the aim of the present work was to examine the presence and nature of PRV-L sequences in a range of plants and to characterize their structure and evolution in selected taxa.

Three main techniques were used to reach the described aims; PCR, Southern hybridization and *in situ* hybridization. These techniques complement each other and have been widely used alone or together to show the presence of and characterize families of repeat sequences from plants and animals (Kubis *et al.*, 1998; Schmidt, 1999; Flavell *et al.*, 1992b; Hirochika and Hirochika, 1993; Miller *et al.*, 1999; Suoniemi *et al.*, 1997). PCR is convenient and fast method to obtain pieces of genomic DNA flanked by short primers of known sequence. Sequencing the product of one or several PCR runs and database comparison will show the diversity of the obtained sequences and relatedness to already published sequences. Southern hybridization of sequences to genomic DNA digests adds to the information by showing the presence in nuclear DNA and to some extent copy number. *In situ* hybridization of labelled sequences to metaphase chromosomes is able to show the actual location of repeat sequences on the chromosomes and thereby also give a good estimate of the copy number (Brandes *et al.*, 1997; Harper *et al.*, 1999).

More specifically, I aimed to isolate genome integrated PRV-L sequences from *Solanum tuberosum* 'Desiree' and analyse their internal structure and organization in the potato genome by Southern and metaphase *in situ* hybridization. I then aimed to investigate the presence and relationship of PRV-L sequences in a range of species from Solanaceae and representatives spanning major groups of the plant kingdom. Further I aimed to analyse the contribution of PRV-L sequences to the genome of plants and their organization by Southern and *in situ* hybridization.

Chapter II: Materials and Methods

2.1 Plant material

Most plants were collected or acquired from commercial sources in England in the area of Norwich (Norfolk) and Leicester (East Midlands) according to table 2.1. Some plants were kept and grown in a greenhouse while others were harvested at the place of growth.

2.2 DNA isolation

Young leaves or fresh green parts of plants were taken for isolation of DNA. The method of Gawel and Jarret (1991) was used with some modifications.

1.5-2 g of the plant material was cut into smaller pieces and packed in silver foil and immersed in a container of liquid nitrogen to assist with the grinding and prevent enzymatic degradation. For each plant variety, a mortar and pestle were cooled by filling with liquid nitrogen and allowing it to evaporate. The plant material was added carefully to the mortar with extra liquid nitrogen and ground quickly to a fine powder. A piece of thick plastic film was cut to size (6x4 cm), the end immersed in liquid nitrogen and used to transfer the plant powder into a 50 ml polypropylene centrifuge tube which was placed in the freezer (-20°C). After all samples had been processed, 20 ml of extraction buffer heated to 65°C was added quickly together with 2 µl of mercaptoethanol (n.b. some people say the mercaptoethanol has to be added to the buffer before coming into contact with the plant material) and the tubes shaken to suspend all the powder. The tubes were incubated for 30 min at 65°C with occasional shaking. 15 ml of chloroform:isoamyl alcohol (24:1) was added and the tubes placed on a rotator for 15 min in the cold room (although the extraction can be done at rt). The extract was centrifuged at 3800-5000 rpm (depending on the capacity of the centrifuge, Jouan C412, 3800 rpm ~2906g) for 10 min at rt. The contents of the tube were now in three parts (aqueous, leaf debris, organic) and the top aqueous supernatant was pipetted into a new 50 ml tube, sometimes using a piece of Miracloth or other dense cloth placed in the opening to filter any plant debris. An equal volume of ice-cold isopropanol was added, usually between 10-15 ml, the tube inverted slowly a few times until the DNA precipitated, and it was then placed on ice or 15 min. The tubes were then centrifuged at 3000 rpm (Jouan C412 ~1811g for 10

Table 2.1. Plant material. The ploidy level and chromosome number, nuclear DNA content (C-value) and origin. Reference for ploidy and chromosome number (refs. mostly taken from "Index of chromosome numbers" Missouri Botanical Garden (also on http://mobot.mobot.org/W3T/Search/ipcn.html): **A**. Zhou Y.-L. *et al.* (1993), Chenia 1:37-42, **B**. Danilkiv & Lobachevskaya (1988) Ukrajins'k. Bot. Zurn 45: 52-54, **C**. Löve & Löve (1976) Taxon 25: 483-500; Kurita (1977) Chromosome Information Service 23: 4-6, **D**. Pinter (1995) Fern Gaz. 15: 25-40; Ormonde & Queiros (1995) Vl. Lagascalia 18(1):63-70, **E**. Druskovic & Lovka (1995) IOPB Chromosome data Newsletter Int. Organ. Pl. Biosyst. (Zurich) 9? 24: 15-19, **F**. Wentworth *et al.* (1991) Watsonia 18: 415-417; 1991, Philosophical Transactions of the Royal Society 334: 309-345, **G**. Singh. R-N. (1992) Cytologia 57: 267-271; Bennet & Smith (1976) Phil. Trans. Royal. Trans B 274: 227-274, **H**. Berg & Greilhuber (1993) Plant Syst. Evol. 185: 259-273, **I**. Wu, H-M *et al.* (1993) Acta Genet. Sin. 20: 50-58, **J**. Tatemichi (1990) Illustrated book of the genus *Nicotiana*, Japan Tobacco Inc. Toyoda, **K**. Wigh (1972) Lindbergia. References for C-value, **1**. Bennett & Leitch, (1995), **2**. Bennett & Leitch, (1997), **3**. Bennett et al. (2000), **4**. Murray, (1998), **5**. Nasu, (1997), **6**. Temsch et al. (1998), **7**. Leitch et al. (1998), **8**. Voglmayr, (2000), **9**. Bennett & Leitch, (2001), 10. Leitch *et al.* (2001), 11. Bennett and Smith (1991).

Taxa	Ploidy	1C-value, pg	Origin
Marchantiopsida			
Marchantia polymorpha	Gam=9 [A]	0.32 [5]	Notcutts Garden Centre, Norwich, UK
Bryopsida			
Leptobryum pyriforme	Gam=20 [B]	0.45 [6]	John Innes, Norwich, UK
Scleropodium purum	Gam=11 [K]	0.34 [8]	Kelling Heath Caravan Park, North Norfolk, UK
Lycopodiopsida			
Selaginella kraussiana	2x = 20	0.06 [9]	University of Leicester Botanical Garden, UK
Equisetopsida			
Equisetum arvensis	Spor=216 [C]	~12 [9]	Dereham, Norfolk, UK
Polypodiopsida			
Polysticum setiferum	?x = 82 [D]	Dryopteris ~4.5 [9]	Notcutts Garden Centre, Norwich, UK
'Gymnospermae'			
Ginkgo biloba	2x = 24	9.9 [4]	Gatersleben, Germany
Pinus pinaster	2x = 24	24.4 [4]	unknown
Gnetum gnemon	2x = 44	3.9 [10]	University of Leicester Botanical Garden, UK
Angiospermae			
Dicotyledoneae			
Nuphar lutea	2n = 34	est. 0.83 [7]	River Yare, UEA, Norwich, UK
Arabidobsis thaliana 'Columbia 0'	2x = 10	0.2 [3]	John Innes, Norwich, UK
Brassica napus 'Falcon'			
Pisum sativum	2x = 14	4.4 [2]	John Innes, Norwich, UK; Biology dept. Leicester University, UK
Atropa bella-donna	6x = 72 [F]	2 [11]	Newarke Houses Museum garden, Leicester, UK

Solanum tuberosum 'Desiree'	4x = 48	1.8 ^{3}	Mousehold Garden Centre, Norwich, UK; Craighill Nurseries, Leicester, UK
Solanum crispum			University of Leicester Botanical Garden, UK
Lycopersicon esculentum	2x=24 ^{I}	1.0 ^{3}	Craighill Nurseries, Leicester, UK
Cyphomandra crassicaulis	2x = 24 ^{I}		University of Leicester Botanical Garden, UK
Brugmansia x candida			University of Leicester Botanical Garden, UK
Cestrum aurantiacum	2x=16 ^{H}		University of Leicester Botanical Garden, UK
Nicotiana tabacum 'SR1'	4x=48	5.8 ^{1}	Biology dept. Leicester University, UK
Nicotiana glutinosa	2x = 24	3.9 ^{1}	John Innes, Norwich, UK
Nicotiana clevelandii	?x = 48 ^{J}	4.0 ^{1}	John Innes, Norwich, UK
Petunia hybrida 'Fancy Pants'	?x = 14 ^{G}	1.7	University of Leicester, UK
Brunfelsia pauciflora			Royal Veterinarian and Agricultural University, Denmark
Monocotyledoneae			
Nerine bowdenii		est. 20.5 ^{7}	John Innes, Norwich, UK
Leucojum aestivum	2x=22 ^{E}	32.0 ^{1}	Earlham House, Norwich, UK
Crocus tommasianus			Norwich churchyard
Musa balbasiana 'Butohan' (BB)	2x = 22	0.5 ^{3}	John Innes, Norwich, UK; KUL, Belgium
Musa x *paradisiaca 'Obino l'Ewai' (AAB)*	3x = 33	0.85 ^{3}	John Innes, Norwich, UK; KUL, Belgium
Oryza sativa	2x=24	0.5 ^{3}	John Innes, Norwich, UK
Zea mays	2x=20	2.7 ^{2}	John Innes, Norwich, UK
Triticum aestivum 'Chinese Spring'	6x = 42	17.8 ^{1}	John Innes, Norwich, UK
Hordeum vulgare 'Sultan'	2x=14	5.5 ^{3}	John Innes, Norwich, UK

min, excess liquid was poured off carefully, followed by a short spin and removal of remaining liquids with a pipette taking care not to disturb the DNA. The DNA pellet was then partly dissolved in TE buffer (250-500 µl) before adding RNase (to a final volume of 10µg/ml) and heating at 65°C for 30 min to remove RNA and completely re-dissolve the DNA. A second precipitation of the DNA was performed by adding 1/10 volume of 3M NaOAc (pH 5.2) and 2 volumes of 96% EtOH. Glass Pasteur pipettes were sealed at the end over a gas flame and used to fish out the DNA, which was then briefly rinsed in 70% EtOH, air dried and dissolved in purified H_2O purchased from Sigma ("Sigma water").

*Extraction buffer 250 ml: 5g CTAB (Cetyltrimethylammonium bromide) (2%); 25 ml 1M Tris-HCl pH 8 (100 mM); 70 ml 5M or 20.5 g NaCl (1.4 M); 10 ml 0.5M EDTA (20 mM), heat to dissolve, H_2O to 250 ml.

*TE buffer 50 ml: 500 µl 1M Tris-HCl pH 8 (10mM); 200 µl 0.25M EDTA pH 8 (1mM) H2O to 50 ml.

*Sodium Acetate (NaOAc) 50 ml: 20.42g of sodium acetate·$3H_2O$ in 40 ml H_2O adjust pH to 5.2 with glacial acetic acid. Autoclave.

2.3 DNA purification

When genomic DNA from the CTAB extraction above was not pure enough, usually because it was found to be sticky or grey/brown in colour, a small portion was purified through a clean-up system to remove recalcitrant secondary products before PCR or other applications. The Promega's Wizard® DNA Clean-Up System was used following the manufacturer's instructions with minor changes. The DNA was re-dissolved in minimum 50 µl of water. A minicolumn (supplied) for each sample was placed in a rack with a syringe attached to the top with the plunger removed. 1ml of resin (supplied) was added and the DNA was slowly added, making sure it fully mixed in. The plunger was inserted and slowly pressed down. The column was then washed by adding 2 ml of 80% isopropanol and pushing it slowly through the column with the plunger. The minicolumn was transferred to a 1.5 ml micro centrifuge tube and centrifuged in a small bench-top centrifuge at maximum speed (c. 10,000 g) for two minutes. The column was transferred to a new tube and 30 - 50 µl of warm (65-70°C) water was applied and the column was incubated for 1 minute at rt. The DNA was diluted in 30 µl (minimum amount) when it was important to keep a high DNA concentration for instance of labelled probes or if the initial DNA concentration was low. The column was centrifuged again for 20 sec at max speed. This purification method was also used to remove excess nucleotides, salts and enzymes following labelling of DNA probes.

2.4 Estimation of DNA concentration

The concentration of DNA, whether genomic DNA, Miniprep DNA, PCR fragments or gel-recovered PCR products for various proposes, was estimated by running on a gel together with a ladder containing bands of known concentration. The DNA was run in a 0.8% agarose gel (0.8 g in 100 ml TAE-buffer) with 3 µl of ethidium bromide (10mg/ml) (EtBr) at 80-100 V for ca. 45 min. The gel was photographed on a UV table and the ethidium bromide fluorescence of individual bands was then compared to the emission of a Lamda (λ) DNA-*Hin*dIII digest ladder with known concentration. (In 0.5 µg of λ ladder the 23130bp band is 225 ng, 9416bp 92.5 ng, 6557 bp 65 ng, 4362bp 45 ng, 2322bp 22.5 ng, 2027 bp 20 ng, 564 bp 5 ng). The λ ladder needs to be heated to 60-65°C for 2 min before loading as two of the fragments (4 and 23 kb) stick together via cohesive ends.

Genomic DNA and miniprep DNA was gel electrophoresed undiluted (1 µl) together with 1 µl or 5 µl of a 1/10 dilution for comparison (made up to a volume of ca. 10 µl with loading buffer and/or H_2O) as these products are often stronger than the control ladder. For gel recovered DNA and PCR products, 5-10 µl respectively were loaded.

*TAE-buffer 500ml 50X stock: 121g of Tris-base; 28.6 ml glacial acetic acid; 50 ml of 0.5M EDTA (pH 8). Working solution 1X.

*Loading buffer: 6X stock: 0.25% Bromophenol blue, 0.25% Xylene cyanol FF, 50% glycerol in H_2O.

2.5 Digest of DNA

For a quick test of DNA quality, 1 µl of DNA was digested in 1µl appropriate buffer (1/10), 1µl of a restriction enzyme (1/10, 10 Units/µl) and dH_2O to 10 µl by incubating at 37°C or other required temperature for a 3 hours. Digests of genomic

DNA for Southern transfer contained more DNA and required longer incubation times to ensure complete digestion. The digest was made up in a volume corresponding to x10 of the DNA volume (5-10 μg) including 1/10 buffer, 1/20 enzyme and dH_2O for the rest. The samples were incubated in a water bath at 37°C (or other required temperature) overnight.

2.6 Precipitation of DNA

A way of concentrating and cleaning DNA (other than the DNA clean-up system) was by ethanol precipitation. To the digested DNA or other product was added 1/10 volume of 3M NaOAc (sodium acetate) and 2.5 volumes of 96% EtOH and placed in the freezer for some hours to overnight. Then the tubes were spun for 30 min at high speed, the supernatant discarded carefully and 0.5-1 ml of 70% EtOH added. The tubes were again spun for 30 min and the supernatant discarded; an extra short spin made it possible to remove excess liquid with a pipette. The DNA pellet was dried for 30 min (rt or 37°C incubator) and dissolved in water (10-30 μl) and, when appropriate, loading buffer.

2.7 Polymerase chain reaction (PCR)

Several primer pairs based on a PRV-L sequence from *Nicotiana tabacum* (Jakowitsch *et al.*, 1999) were selected as 18-23 mers and ordered from Gibco BRL, Life technologies. TPV890L-1909R, TPV3462L-4579R, TPV6241L-7041R, or Sigma-Genosys, TPV6118L, TPV6339L, TPV7037R, TPV7072R, numbers refer to the assembled virus-like sequence (database ID, NTA238747) (figure 2.1 and table 2.2). The following degenerate primers for *copia* and *gypsy* retrotransposon RT domain was used Ty1-1 and Ty1-2 (Flavell *et al.*, 1992a), GyRT1 and GyRT4 (Friesen *et al.*, 2001). The conditions used for the PCR reactions were based on one cycle of 93 °C for 5 min, 35 cycles of 94°C for 30 sec, annealing temperature (X) °C for 30 sec, 72°C for 90 sec and finally one cycle of 72°C for 5 min., with X varying between 45 and 56°C (see table 2.2).

To eliminate contamination problems, PCRs were set up in a flow hood carefully rinsed with ethanol and DNAaway (Fisher Scientific), pipettes and rack were rinsed in the same way, and one pipette was used for all the PCR components except DNA which was taken with a separate pipette. A control with primers but

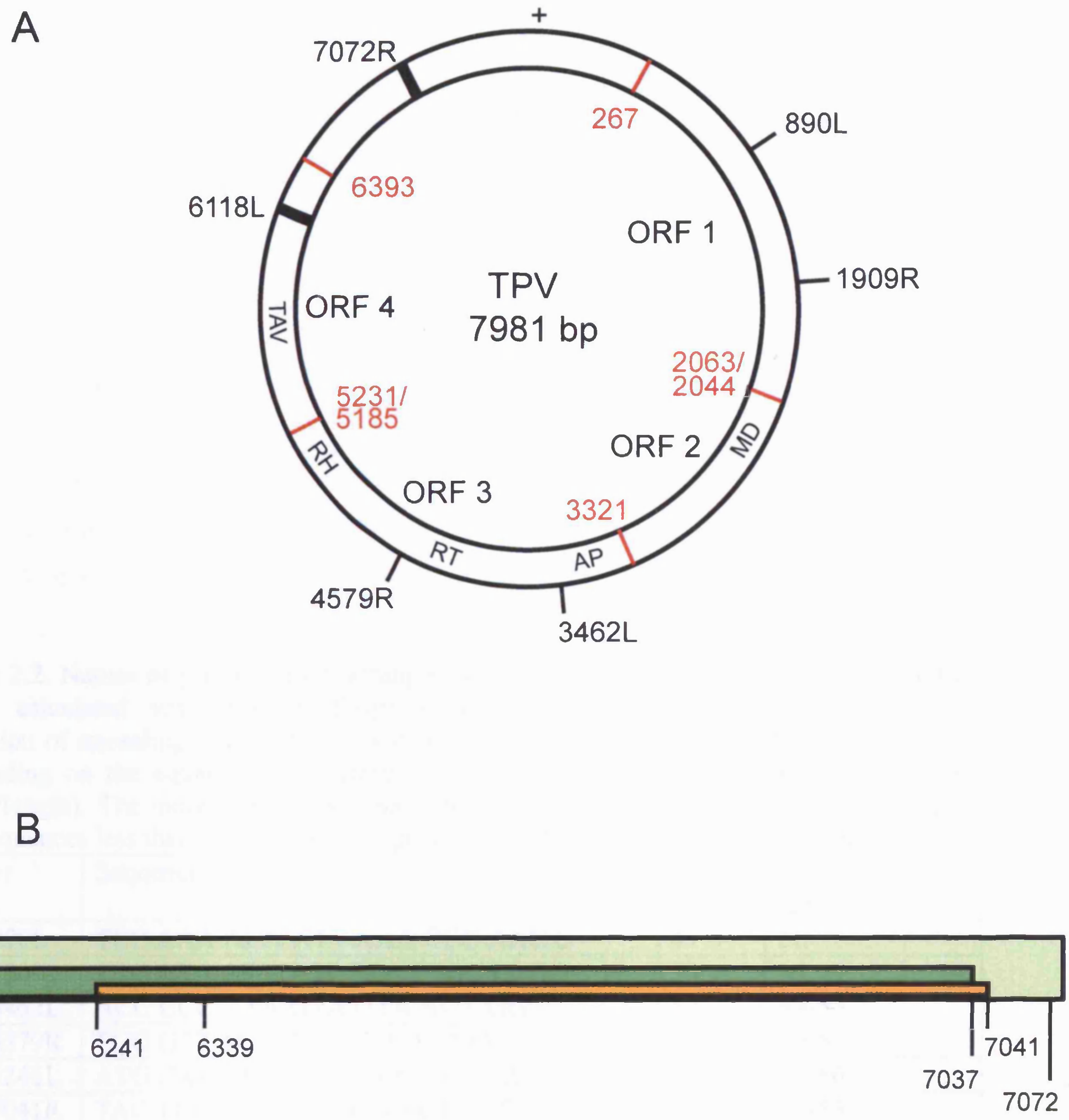


Figure 2.1. A) Schematic diagram of tobacco pararetrovirus (TPV) as a circular dsDNA molecule, showing the four open reading frames (ORF) and the transcription start site (+), numbers indicate nucleotide positions of ORFs (red, inside circle) and primers (outside circle). The thick dark bands in the ring at the end of ORF four indicate the area for PCR amplification where the position of the individual primers are shown in detail in B). MD movement domain, AP aspartic protease, RT reverse transcriptase, RH RNase H. The numbers in A) and B) correspond to the primers in table 2.2. and is the nucleotide positions of the primers relative to TPV (Jakowitsch et al., 1999). The light green bar is 935 bp, the dark green 899 bp and the orange is 800 bp.

Table 2.2. Names of primers used, arranged in pairs, their sequence, melting temperature (Tm) calculated with JustBio (http://www.justbio.com/oligocalc/index.php) and the variation of annealing temperatures used in PCR reactions. The calculated Tm will wary depending on the equation used. Here is used the equation: Tm = 64.9 + (41x(G+C-16.4)/length). The more simple (Wallace rule): 2x(A+T) + 4x(C+G) is usually only used for sequences less than 14 bases and originally made for membrane hybridizations.

Primer name	Sequence 5'-3'	Tm °C	PCR, annealing temp. °C
TPV890L	TCC AAA ACA ATT AAA CCC AAC C	49	56
TPV1909R	TCG GTA ATT TGT TTC TTC TTT GG	50	56
TPV3462L	ACC CCC ATA GTA GTA ACA GG	52	49-56
TPV4579R	TGG GTT TGC ATT TTT ATT CC	46	49-56
TPV6241L	ATG CAA CCA ACT ACC AGA GC	52	49-56
TPV7041R	TAC TCC CCT AAA CGG CTT CC	54	49-56
TPV6118L	GGA AAT GCT ATC TGG AGC	48	49-52
TPV7072R	TCA TTT ACC CTA ACA GCG CC	52	49-52
TPV6339L	ATG TTC AAA GTG CAA CGG AG	50	49-52
TPV7037R	CCC CTA AAC GGC TTC CTT GCC	58	49-52
28S-1R	GCT ATC CTG AGG GAA ACT TCG G	57	any
28S-1F	CCG TCT TGA AAC ACG GAC CAA GGA G	61	any
Tyl-1 L	ACN GCN TTY YTN CAY GG	40-54	43-44
Tyl-2 R	ARC ATR TCR TCN ACR TA	35-47	43-44
GyRT1 L	MRN ATG TGY GTN GAY TAY MG	42-58	43-44
GyRT4 R	RCA YTT NSW NAR YTT NGC R	38-55	43-44
M13 L	TGT AAA ACG ACG GCC AGT	48	
M13 R	CAG GAA ACA GCT ATG ACC	48	
T7	AAT ACG ACT CAC TAT AGG G	47	

without DNA was run. Positive control of the PCR and DNA was performed with primers amplifying 28S ribosomal DNA (rDNA).

PCRs were set up in a total volume of 50 µl: 5 µl 10x reaction buffer (1x); 1.5 µl $MgCl_2$ (1.5mM); 4 µl dNTPs (10mM); 2 µl of each primer (20 pmol); x µl DNA (10-40ng), Sigma water to a total of 50 µl, 0.25 µl *Taq* polymerase (1.25U). If the primers were the same in all reactions a master mix was made with all reagents except the DNA, mixed thoroughly and distributed to 0.2 ml thin walled tubes, closing lids, and the individual DNAs were added quickly to each tube with another pipette. Reactions were kept on ice until the thermocycler had reached ca. 80°C and then quickly placed in the wells of the machine (T-Gradient thermocycler or Perkin Elmer). BIOTAQ^{TM} DNA polymerase from Bioline (5u/µl) was used including their 10x reaction buffer (160mM $(NH_4)_2SO_4$, 670mM Tris-HCL (pH 8.8 at 25°C), 0.1% Tween-20) and $MgCl_2$ (50mM) solution. The products were size separated in a gel by electrophoresis as under "Estimation of DNA concentration".

*Primer stock made from manufactured primers: 100 pmol/µl, working dilution 1:10 (10 pmol/µl). Amount of primer in nmol x 10 = amount of dH_2O to add to get 100 pmol/µl.

*dNTP solution (2.5 mM): 1/40 of 100mM dNTPs, e.g. 5 µl of each dATP, dCTP, dTTP and dGTP in 180 µl dH_2O. dNTP set, PCR grade from Roche.

2.8 Recovery of DNA from gels

PCR products (or other DNA fragments) were extracted as a band from a gel to exclude other DNAs. The whole (50 µl) of the PCR reaction was loaded onto a gel (1.2g agarose in 120 ml TAE buffer) including 6-8 µl loading buffer. Each band was cut narrowly from the gel with a knife blade and placed in a 1.5 ml micro centrifuge tube. The DNA was extracted from the agarose with Nucleospin Extract kit from Macherey-Nagel following the manufacture's instructions. To test the concentration of the recovered and diluted DNA, 5 µl DNA was loaded on a gel together with 5 µl loading buffer.

2.9 Competent cells

Preparation of electro-competent *E. coli*: To minimize contamination, bottles and tubes were washed with 1% SDS and rinsed thoroughly with dH_2O and autoclaved filled with dH_2O. All steps after the cell culture has been cooled were performed as quickly as possible, usually on ice. The starting culture material was a LB plate streaked with a toothpick dipped in a glycerol stock of DH5α cell line grown at 37°C. One colony was transferred to a flask with 10 ml of LB-g medium and incubated in a shaker (200 rpm, Gallenkamp) overnight at 37°C. 5 ml of that culture was transferred to a flask of 500 ml pre-warmed 2xYT medium. After some hours the optical density of 1 ml of the culture was measured at 600 nm. A blank of 1 ml of plain 2xYT medium was used. When the cell culture reached an OD of 0.5-0.6 the flask was packed in ice in a cold room (4°C) and cooled for 30 min. The culture was divided into 2 x 250 ml centrifuge tubes and spun at 4500 rpm (Sorvall RC-5B) for 15 min at 4°C. The supernatant was then carefully poured off. A small amount of ice cold HEPES buffer was added and the tubes swirled and knocked carefully to re-suspend the cells, then the tubes were filled with ice cold HEPES buffer and centrifuged as above (4500 rpm for 15 min, Sorvall RC-5B). The cells were re-suspended first in a small amount of HEPES buffer and then a larger amount and set to rest for 20 min in ice in a cold room (4°C). Then the cells were centrifuged as above again, supernatant poured off and pellet re-suspended slowly in 100 ml ice-cold HEPES buffer and centrifuged as above. The supernatant was poured off and cells suspended slowly in 30 ml ice-cold 10 % glycerol and centrifuged at 4500 rpm (Sorvall RC-5B) for 10 min at 4°C. The supernatant was poured off and the pellet in each tube re-suspended in 1.5 ml ice-cold 10% glycerol and placed on ice. In a cold room (4°C) the cells were aliquoted (50 μl) with cold tips into cold micro centrifuge tubes, fast frozen in liquid nitrogen and stored in a -80°C freezer.

*LB-g medium 1 litre: 10g Tryptone; 5g yeast extract; 10g NaCl; dH_2O to 900 ml, shake to dissolve, adjust pH to 7.0 with 5N NaOH, fill to 1 litre, autoclave.

*2xYT medium 1 litre: 16g tryptone; 10 g yeast extract; 5g NaCl; dH_2O to 900 ml, shake to dissolve, adjust pH to 7.0 with 5N NaOH, fill to 1litre, autoclave.

*HEPES ($C_8H_{18}N_2O_4S$) buffer: 119 mg/0.5 litre dissolved in ddH_2O, autoclaved.

*10% Glycerol: 10 ml glycerol in 90 ml of ddH_2O.

2.10 Ligation

For ligation of a PCR fragment, usually a band recovered from a gel, a kit "pGEM®-T Easy Vector system I" from Promega was used. A reaction had a total volume of 10 µl in 0.5 ml tubes: 5 µl of 2x Rapid Ligation Buffer (1x) (provided); 1 µl of plasmid vector (1ng) (provided); x µl PCR product; y µl dH_2O and 1 µl T4 DNA ligase. The amount of DNA added was dependent on the size of the PCR product and increased with increased size following a formula and assuming that a 3:1 ratio of insert to vector is optimal: ((Amount of vector (50ng) x size of insert in kb) divided with size of vector 3.0kb) x 3/1 for the ratio = amount of insert in ng. Example with an 800 bp insert: ((50ng x0.8kb)/3) x3/1=40 ng. The reaction was incubated for 1 hour at rt and then overnight at 4°C for maximum number of transformants.

*2x Rapid Ligation Buffer: 60mM Tris-HCl (pH 7.8); 20mM $MgCl_2$; 20mM DTT; 2mM ATP; 10% polyethylene glycol.

2.11 Transformation

The ligation reaction was then transformed into the prepared competent cells by electroporation or heat shock. 1-2 µl of ligation mix was added to a thawed tube with 50 µl competent cells and mixed briefly by tapping.

Electroporation. The content of the tube with competent cells and ligation mixture was transferred to a cooled transformation cuvette ensuring that no air bubbles were present and placed in the holder of the electroporator (Gene Pulser II). Current was applied at 2.5 V for a few seconds with the following settings: "low range" between 100-200, "high range" on lowest setting (500) and "capacity" on 25. After electro-shock the cells were quickly transferred (with a plugged glass pastette) to a 1.5 ml micro centrifuge tube containing 1 ml SOC medium and placed in a shaker at 37°C for 1 to 1½ hour. The tube was then spun at low speed (3500 rpm, 1164g) in a micro centrifuge to bring all cells to the bottom and the pellet was re-suspended in 200 µl of SOC medium, and plated on two plates with selection medium and was placed in a 37°C incubator overnight.

Heat Shock. The cell ligation mixture was incubated for 30 min on ice. The tube was then placed in a 42°C water bath for 90 sec (no shaking) and then immediately on ice for 5-10 min. 500 µl LB medium was added to the tube and the culture incubated in a shaker for 1 hour at 37°C. 150 µl of the culture was plated on each plate containing selection medium and placed in a 37°C incubator overnight.

Cloning of a DNA fragment into this particular vector disrupts the coding sequence of β-galactosidase, and successfully transformed bacteria will usually give rise to white colonies. In the flow hood, a number of white colonies were picked with a toothpick and each placed in a small flask with 10 ml LB medium and 4 µl ampicillin (100mg/ml) giving a concentration of 40µg/ml. The flasks were placed in an incubator at 37°C overnight. The plasmids were isolated from the overnight cultures by a mini plasmid preparation kit from Promega (Wizard® Plus Minipreps DNA Purification System) following the manufacturer's instructions, using 4.5 ml of the overnight culture instead of 1-3 ml.

The presence of the insert in the plasmids were confirmed before plasmid isolation by PCR with 4 µl of the overnight culture using M13 primers or after plasmid isolation by restriction digestion of plasmid DNA using an appropriate restriction enzyme which cuts out the insert. Here it was *Eco*RI.

*SOC medium 100ml: 2.0g tryptone; 0.5g yeast extract; 1 ml 1M NaCl; 0.25ml 1M KCl in 97 ml of dH_2O, stir to dissolve, autoclave. Add 20mM 2M Mg^{2+} stock and 20mM 2M glucose, fill to 100ml, filter sterilize. pH 7.

*2M Mg^{2+} stock 100ml: 20.33g $MgCl_2 \cdot 6H_2O$; 24.65g $MgSO_4 \cdot 7H_2O$; dH_2O to 100ml, filter sterilize.

*2M glucose 100ml: 36g glucose, dH_2O to 100ml, filter sterilize.

*Agar plates with selection: LB medium with 15g agar per litre, autoclaved, and aliquoted in flasks of 200ml portions. For making plates a flask was heated in a microwave oven at a low setting (level 3 for 12 min) until all the medium was liquid, then allowed to cool to 60°C before adding (per 200ml) 200 µl ampicillin (100µg/ml); 400 µl X-gal (40µg/ml) and 500 µl IPTG (0.5mM) in a flow hood. Swirl slightly and pour the plates quickly. This amount usually gave 8-9 plates.

*Ampicillin 100mg/ml: 1g ampicillin in 10 ml sterile water, aliquot, store at -20 °C.

*X-gal 20mg/ml: 320mg 5-bromo-4-cloro-3-indolyl β-D-galactoside; 16 ml N,N'dimethyl-formamide (DMF), aliquot and wrap in aluminium foil, store at -20°C.

*IPTG 200mM: 476 mg isopropyl- β-D-thiogalactopyranoside, 10 ml sterile water, aliquot, store at –20 °C.

2.12 Sequencing

For sequencing, either a solution of recovered plasmids or PCR products were supplied to the Protein and Nucleic Acid Chemistry Laboratory (PNACL, Leicester University). For a PCR product about 110 ng was supplied in 8 µl Sigma water in 0.5 ml micro centrifuge tubes. For plasmids a larger amount of DNA was supplied, about 400 ng. For sequencing of plasmids standard primers supplied by the sequencing department e.g. M13 forward and reverse or T7 were used. For sequencing of un-cloned PCR products the original primers had to be supplied at a concentration of 0.8-1.0 pmol/µl in 10 µl of water. From a primer stock of 100pmol/µl this is 1:111.

PNACL carried out cycle sequencing reactions, purified them using DyeEx columns and analysed the products using an ABI 377 automated sequencer. The sequence with the chromatogram was submitted in computer format. The chromatogram was viewed and edited in the PC program Chromas (www.technelysium.com.au/Chromas. html). The sequence was submitted to Fasta3 on the EBI web page for comparison to known sequences (www.ebi.ac.uk).

Previously the cycle sequencing reaction and clean-up step was performed before giving the product to the sequencing service, John Innes Centre, Norwich. The DYEnamicTM ET terminator cycle sequencing kit (Amersham Pharmacia Biotech) was used following the manufactures protocols: sequencing reaction premix (supplied) 4 µl, DNA (0.1-0.2 pmol, 250-500 ng plasmid DNA), primer 2 µl (5 pmol) (one reaction for each of forward and reverse primer), dH_2O to 10 µl. The reagents were mixed by tapping and overlaid by a drop of mineral oil and placed in the thermal cycler (PerkinElmer) programmed as follow: 20-30 cycles of 95°C for 20 sec, 50°C for 15 sec, 60°C for 1 min. To the finished PCR reaction was added 50 µl of dH_2O and it was centrifuged for 1 min before removing the 60 µl from under the oil. For precipitation 6 µl of NaOAc was added (provided with the kit) and 200 µl EtOH. The sample was placed at –20°C overnight and then centrifuged at 4°C for 30 min, the supernatant dissolved in 300 µl 70% EtOH and centrifuged as above. After discarding the supernatant the tubes got a brief spin and remaining liquid taken carefully with a pipette. Pellet was dried at 37°C for 5-10 min. This product was brought to the sequencing service.

2.12.1 Alignment and tree construction

Sequences were aligned in ClustalW from EBI homepage (www.ebi.ac.uk) using default settings (full alignment, matrix iub, gap open 10, gap ext 0.05, gap dist 8) and manipulated to have homologous 5' and 3' ends. Phylograms were made from ClustalX with the Neighbour Joining method using the algorithm of Saitou and Nei (1987) and opened in TreeView (Page, 1996). Bootstrap values, a statistical estimate of the reliability of some groupings, were generated from 1000 trials.

2.13 Southern hybridization

For both non-radioactive and radioactive Southern blotting the gels were treated in the same way after the principle of Maniatis *et al.* (1982). Gel properties and running condition was as follows: 1% agarose in TAE- or TBE buffer plus 5µl ethidium bromide preferable run overnight on a low setting i.e. 30V for 15-20 hours. A photograph was taken of the stained gel with a fluorescent ruler at the side of the gel. The gel was cut to size and immersed in 0.25M HCl (depurination) on a shaker for 10 min. Then there were two further washes for 45 min in denaturation solution and neutralization solution. The gel was washed briefly in transfer solution (10xSSC) before it was placed in the set-up for a traditional upward capillary transfer: a "table" large enough to support the gel was placed in a tray containing transfer buffer (10xSSC), with one piece of moistened 3MM Whatman filter paper over it. The gel was placed on Whatman paper upside down. The membrane (neutral nylon membrane Hybond^{TM}-N, Amersham Pharmasia Biotech) briefly wetted in dH_2O was marked and placed carefully on top followed by three pieces of Whatman paper cut to size and then stacked with 5-10 cm of tissue/towel paper and a weight on top. Pieces of parafilm were placed around the edge of the gel to prevent liquid to bypass the gel. Transfer was carried out overnight (12-24 hours).

*TBE-buffer 1 litre 5X stock: 54g Tris base; 27.5g boric acid; 20 ml 0.5M EDTA (pH 8); H_2O to one litre.

*Depurination solution 1 litre: 21.6 ml 11.6M HCL (0.25M); H_2O to one litre.

*Denaturation solution 500 ml: 43.9g NaCl (1.5M); 10.0g NaOH (0.5M); H_2O to 500 ml.

*Naturalization solution 1 litre: 87.7g NaCl (1.5M); 500ml 1M Tris HCl (0.5M); H_2O to one litre.

*Tris HCl (1M) 1 litre: 121.1g Tris base in 900 ml H_2O + concentrated HCl to pH 7.4.

*Transfer buffer, 10 x SSC: 1:2 of 20xSSC 1 litre: 175.3g NaCl; 88.2g $Na_3C_6H_5O_7.2H_2O$; H_2O to one litre (pH 7 is usually reached without further adjustment), autoclave.

2.13.1 Non-radioactive detection

Hybridization and detection was carried out using Alkphos (Amersham Pharmacia Biotech) using their protocol for hybridization buffer and primary- and secondary wash buffer.

*Hybridization buffer: 60 ml of provided hybridization buffer (12% w/v urea, 2M); 1.76 g NaCl (0.5M); 2.4g blocking reagent (4% w/v) for 4 portions.

*Na-phosphate (1M) 200 ml: 23.99g $NaH_2PO_4xH_2O$, adjust pH to 7.0 with NaOH.

*Primary wash buffer 500 ml: 60 g Urea (2M); 50 ml 1% SDS (0.1%); 25 ml 1M Na-phosphate (50 mM); 4.4g NaCl (150 mM); 0.5 ml 1M $MgCl_2$ (1mM); 1g blocking reagent (0.2%).

*Secondary wash buffer 500 ml 20x stock: 60.5g Tris base (1M); 58.4g NaCl (2M), adjust pH to 4. For working solution dilute 1:20 and add 2ml/litre $MgCl_2$ (2mM).

For a hybridization tube (12.5x3.5 cm) 15 ml of hybridization buffer was used with 120 ng of probe. DNA for probe had to be diluted to 10ng/µl before labelling, 12 µl labelled DNA was used. Blots were prehybridized for 2-4 hours at 55°C before adding the labelled probe to hybridize overnight at 55°C. After washing, the blot was placed on a piece of cling film in the dark room and the detection reagent (CDP-star) was added. The membrane was drained briefly and quickly wrapped in cling film, avoiding air bubbles and folds, and placed in the cassette secured with tape. A film (Fuji Super RX) was placed on top in complete darkness. A trial exposure was carried out, usually for 2-3 hours.

For reprobing the blot was washed following the protocol: incubation in 0.5% SDS at 60°C for 1 hour and rinsed in 100 mM Tris pH 8 for 5 min at rt, stored wet wrapped in cling film.

2.13.2 Radioactive detection

Labelling of probe with Megaprime DNA labelling system from Amersham Pharmacia Biotech: about 25 ng of DNA; 5 µl nonamer primers (supplied); H_2O to 33 µl. The mixture was denatured for 5 min in a boiling water bath, then cooled on ice. In the radioactivity lab add 10 µl labelling buffer; 2 µl enzyme; 5 µl α-32P-dCTP (1.85 MBq, PerkinElmer) to have a final volume of 50 µl. Place in a 37°C water bath for 10-15 min at least before use; hybridize overnight at 65 °C.

Low stringency hybridization was carried out at 42°C and 5xSSPE for prehybridization and 0.5xSSC/1% SDS for hybridization. Probes were labelled with a random primer DNA labelling system kit from Gibco BRL Life Technologies.

Blots were placed in a hybridization tube (roller bottle) together with 20 ml prehybridization solution at 65°C and incubated for several hours. The probe was prepared and added to the blot together with 2.5 ml hybridization solution after the prehybridization solution had been poured off. The blot hybridized overnight at 65°C.

*Prehybridization solution 40 ml (2 portions): 2 ml 1M Tris-HCl pH 8 (50mM); 0.8 ml 0.5M EDTA (10mM); 10ml 20xSSC (5xSSC); 0.8ml 50x Denhardts (1x); 0.8ml 10% SDS (0.2%); 25.2 ml dH_2O; 0.4 ml Salmon sperm DNA 10mg/ml (have to boiled for 10 min and placed on ice before adding to 65 °C pre-warmed buffer.

*Hybridization solution 5 ml (2 portions): 1 ml 50% Dextran sulphate (10%); 0.25ml 1M Tris-HCl pH 8 (50mM); 0.1ml 0.5M EDTA pH 8 (10mM); 1.25ml 20xSSC (5xSSC); 0.1ml 50x Denhards (1x); 0.1 ml 10% SDS (0.2%); 2.2 ml H_2O; 0.05ml Salmon sperm DNA 10mg/ml (have to boiled for 10 min and placed on ice before adding to 65°C pre-warmed solution.

*50x Denhardts: 1% (w/v) Ficoll400; 1% (w/v) polyvinylpyrrolidone; 1%(w/v) bovine serum albumin in 100ml H_2O.

*50% Dextran sulfate: 5g in10 ml H_2O, filter sterilize and aliquot. Keep at –20°C.

*Wash buffer 100ml 2x2 washes: 10ml 20xSSC (2x); 5ml 10% SDS (0.5%); 85 ml H_2O.

2.14 Preparations for *in situ* hybridization

2.14.1 Harvest of root tips

Young, actively growing white root tips were picked in the morning around 10 to 13 o'clock depending on the time of year (e.g. earlier in the summer) and placed in vials containing 2-3 ml 2 mM 8-hydroxyquinoline (0.29 g/litre) for either 24 hours at 4°C or 1-2 hours at rt and then 1-2 hours in the fridge/cold room (4°C). The roots were then fixed in 3:1 ethanol:acetic acid and left at rt for about 24 hours, after which they were stored at -20°C for up to 2 months.

2.14.2 Pre-treatment of roots

Root tips for *in situ* hybridization preparations had to be enzyme treated before squashing to soften the cell walls. First they were washed in 1x enzyme buffer for 2x 15 min with change of buffer and then transferred to 37°C warm enzyme solution and incubated for 20 min to 90 min depending on root thickness and age of enzyme solution. Transferring the roots to cold (fridge) enzyme buffer at rt largely stopped the digestion.

*Enzyme buffer 10x: 40% 100 mM citric acid $C_6H_8O_7.H_2O$ (2.1 g in 100 ml dH_2O) and 60% 100mM trisodium citrate $Na_3C_6H_5O_7.2H_2O$ (2.9g in 100 ml dH_2O), store at 4°C.

*Enzyme solution: 0.8g Calbiochem cellulase; 0.2g cellulase "Onuzuka", 1.9g Pectinase; dH_2O to 50 ml.

2.14.3 Chromosome squashes

Each enzyme-treated root was transferred to one end of a cleaned frosted glass slide with a plastic pastette, excess enzyme buffer was removed and a drop of 60% acetic acid applied to briefly soak the root. The very tip of the root was then transferred to a small drop of 60% acetic acid in the middle of the slide using tungsten needles. The root cap and outer layers were removed from the inner softer cells containing the metaphase cells. The remaining cells were spread to one cell thickness and an 18x18mm glass cover slip applied. Light pressure was applied to the cover slip and if needed briefly heated over a spirit flame to spread the cytoplasm. After viewing under phase contrast on a light microscope the slide was placed on a metal tray on dry ice for about five minutes and the cover slip was flicked off with a razor blade and the slide air-dried. After some 3-5 days when the slides had dried completely, they were stored at 4°C.

2.14.4 Labelling of probes

Random primer labelling was performed with a kit for radioactive labelling where the radiolabelled dNTP was substituted with biotin or digoxigenin labelled dUTP: For probes 1-5 kb in length (PCR products and inserts from plasmids), labelling was performed with the Random Primers DNA Labeling System (Gibco BRL Life Technologies). The protocol of the manufacturer is aimed at radioactive probes so the procedure was altered somewhat. Approximately 100 ng of DNA was dissolved in 26 μl of dH_2O and the DNA denatured by placing the tube in boiling water for 5 min. and rapidly cool on ice. The supplied reagents were added: dCTP 2 μl; dATP 2 μl; dGTP 2 μl; dTTP 1 μl; Random Primer Buffer Mixture 15 μl and either 1 μl biotin-16-dUTP (stock 25 nmol) or 1 μl digoxigenin-11-dUTP (stock 25 nmol). Just before incubating at 25°C, 1 μl Klenow Fragment was added. After 1 hour 5 μl of stop buffer was applied and the DNA extracted in 30 μl water with Promega Wizard DNA clean-up Systems described earlier.

PCR labelling: For small DNA fragments up to 800 bp max, the procedure of Schwarzacher and Heslop-Harrison (2000) was followed: 5 μl PCR buffer; 1.5 μl $MgCl_2$; 1.5 μl dNTPs (2mM); 1.5 μl of each primer (M13 Forward and reverse for plasmids); x μl DNA (10-30ng) and H_2O to 49.5 μl were mixed and 0.5 μl *Taq* DNA polymerase was added. The PCR programme was 93°C for 5 min, then 35 cycles of 94°C for 30 sec, 56°C for 30 sec, 72°C for 90 sec, and a final step of 72°C for 5 min. Probes were cleaned with Promega Wizard DNA clean-up Systems or ethanol precipitation.

Nick translation was used for genomic DNA sheared to 2-10 kb or linearized plasmids. The procedure followed Schwarzacher and Heslop-Harrison (2000), 1 μg DNA was diluted to 34 μl with dH_2O, with addition of 3 μl unlabeled nucleotides (dCTP, dGTP, dATP, 0.5mM each); 2 μl labelled dUTP mixture; 1 μl DTT; 5 μl 10 x nick translation buffer. After mixing, 5 μl DNA polymeraseI/DNaseI solution was added and the tube incubated at 15°C for 1½ hour. The reaction was stopped with 3 µl 0.5M EDTA. Probes were cleaned with Promega Wizard DNA clean-up Systems or ethanol precipitation.

Incorporation of label was checked by dot-blot following the method of Schwarzacher and Heslop-Harrison (2000). 1 µl of the probe was spotted on a membrane (as used for Southern). The membrane was soaked in a buffer containing blocking reagent for 30 min before application and incubation with antibody-AP mixture (e.g. anti-dig and anti-bio) for another 30 min. After washing the membrane detection reagent was added and left in the dark for 5-10 min. Alternatively a small amount of the probe was run in a gel and transferred to a membrane to be detected as above.

*Labelled dUTP nucleotide mixture: 2 µl Digoxigenin-11-dUTP; 4 µl TTP; 4 µl H_2O or 4 µl biotin-16-dUTP; 0.8 µl TTP; 3.2 µl H_2O.

*DTT (dithiothreitol) 100mM: 77mg in 5 ml dH_2O.

*Nick translation buffer (10x): 500 mM Tris-HCL pH 7.8; 50mM MgC12; 5 mg/ml BSA

*EDTA 0.5M 1 litre: 186.1 g disodium ethylene diamine tetraacetate·$2H_2O$ in 800 ml H_2O. Add NaOH pellets to pH 8.0. Autoclave.

2.15 *In situ* hybridization

The procedure of *in situ* hybridization followed Schwarzacher and Heslop-Harrison (2000) with minor modifications.

Pre-treatment: chromosome slides were re-fixed in 3:1 ethanol: acetic acid for 10 min then dehydrated in 96% EtOH for 2x 10 min and air-dried. 200 µl RNase (1:80 of stock 10mg/ml) in 2x SSC was applied per slide and covered with a plastic cover slip and incubated for 1h at 37°C in a moist chamber. The slides were then washed in 2x SSC for 2x 5 min and 0.01M HCl for 5 min. Enzyme digestion with pepsin (1:150 - 1:750 from a stock of 1mg/ml in 0.01M HCl) depending on the amount of cytoplasm present, 200 µl were added to each slide covered and incubated for 10 min at 37°C in a moist chamber. Slides were then rinsed in distilled H_2O for 1 min. and 2xSSC for 2x 5 min. Finally the slides were treated with 4% paraformaldehyde for 10 min. and rinsed in 2x SSC for 2x 5 min before dehydration in an alcohol series: 70%, 90% and 96% EtOH for 2 min each and air dried.

*20xSSC (saline sodium citrate) 1 litre: 175.3g NaCl; 88.2g $Na_3C_6H_5O_7 \cdot 2H_2O$ (trisodium citrate); dH_2O to one litre, (pH 7 is usually reached), autoclave.

*Moisture chamber, heated to 37°C: metal box with wet tissue in the bottom and a "rack" of two glass rods held together by rubber tubes.

*Paraformaldehyde 4%: 4g in 80 ml H_2O, heat to 60°C, add strong NaOH to clear, adjust to pH 8 with 1N H_2SO_4, H_2O to 100 ml. Cool down. NB. The solution has no buffer capacity.

2.15.1 Hybridization

A master hybridization mix was made from the chemicals in common (usually all except probe DNA) and divided into individual tubes before adding individual probes. The probe mixture was denatured at 70°C for 10 min in a water bath and cooled for at least 5 min on ice. The probe mixture was added to the slide and covered with a plastic cover slip and transferred to a Hybaid thermal cycle oven for whole chromosome denaturation at 70-78°C with slow drop in temperature to 37°C, the temperature of the overnight incubation (20 hours). For some experiments the slides were transferred to Hy-Pro machine with vibrating effect (AVS setting 3) for overnight hybridization as this new technique might be beneficial for the hybridization efficiency as the shaking gives a larger chance for the probe to be moved to the target chromosome site.

*Probe mixture contained: formamide (15-50%); 20xSSC (1-2x) - together giving a stringency of 54-76%; 50% dextran sulphate (10%); 1µg/µl sonicated salmon sperm (1µg); 10% SDS (0.15%), probe (1-5 µl); H_2O to a total volume of 30 or 40 µl.

*SDS (sodium dodecyl sulfate) 10%: 10 g to be dissolved in 100 ml dH_2O.

*Dextran sulfate 50%: 5 g in 10 ml dH_2O, filter sterilized.

2.15.2 Washing

After the hybridization the slides were washed at 42°C in a shaking water bath, ensuring that the formamide mix was within +/- 0.5°C for accurate stringency control. The wash procedure involved 2x SSC to remove coverslip; 2x SSC for 2 min; formamide mix 2x 5min; 1 to 0.1x SSC for 5 min; depending on stringency required; 2x SSC for 5 min; 4x SSC/Tween for 5 min. Then rt: 4x SSC/Tween for 2x 5 min.

*Formamide mix: 100% formamide is best stored in aliquots at -20°C: 20%: 40 ml formamide; 1 ml 20x SSC, 159 ml H_2O.

*4xSSC/Tween (detection buffer) 1 litre: 200ml 20x SSC; H_2O to nearly one litre; 2 ml Tween 20; H_2O to one litre, mix.

2.15.3 Detection, mounting and image capture

The slides were pre-treated with a 5% BSA block (0.1g in 2ml 4x SSC/Tween), 200µl per slide for 5 min at rt before addition of a detection mixture. The detection mixture depended on the label on the probes but would often be a combination of either Cy3-avidin or Alexa-594-avidin (1/200 from 200 µg/µl) and anti-dig FITC (fluorescein isothiocyanate) (1/75 from 200 µg/ml) in BSA block. 40µl were added to each slide and they were incubated in a moist chamber for 1 hour at 37°C. Excess detection reagent was rinsed off with 4x SSC/Tween20 at 37°C for 3x 5 min. For counter staining of the chromosomes DAPI (4', 6-diamidino-2-phenylindole) was used in a concentration at 4µg/ml. 100µl was added to the slides with a cover slip over and incubated for 10 min in darkness. Slides were then washed briefly with detection buffer and mounted with Citifluor by applying a drop to a large cover slip turning it around and sliding it carefully down over the chromosome area. The cover slip was pressed down with filter paper to remove excess liquid.

Slides were examined on a Zeiss Axioplan fluorescence microscope using x16 and x100 oil immersion lenses. Images were captured with three different filters, appropriate to the fluorescence being detected, with a CCD camera.

Chapter III: Isolation and characterization of pararetrovirus-like sequences from the genome of potato (*Solanum tuberosum* 'Desiree')

3.1 Introduction

The aim of the work in this chapter was to analyse the genome of potato (*Solanum tuberosum* 'Desiree') for the presence of PRV-L sequences. The plan was to design primers from the previously isolated TPV sequence from tobacco (Jakowitsch *et al.*, 1999) and use these in PCR to find homologous sequences in potato genomic DNA. The intention was to use the PCR products to obtain clones and investigate any diversity by sequence comparison and characterize the genome organization and copy number of PRV-L sequences by genomic Southern hybridization and *in situ* hybridization.

Jakowitsch *et al.* (1999) initially found a 2kb fragment in tobacco nuclear DNA with homology to RT of some pararetroviruses. From healthy tobacco plants they prepared λ genomic libraries and probed them with this fragment and got hundreds of positive clones. From sequences of 22 independent insertions they assembled an 8 kb PRV-L genome which, at that time, was a previously unknown pararetrovirus and was called TPV. The individual fragments were 91-98% similar and contained stop codons and frame shifts and they defined the TPV sequence from tobacco as a new class of dispersed repetitive DNA in plants. A year later Lockhart *et al.* (2000) published the finding of a very similar PRV-L sequence in *N. edwardsonii* which they had named TVCV. Unlike Jakowitsch *et al.* (1999) they detected episomal viruses, but only under certain environmental conditions and it was only transmitted by seed. Both studies showed TPV/TVCV hybridization to genomic DNA of the allopolyploids and at least one parent. An unusual feature of TPV and TVCV was a very repetitive region after ORF4 (TAV) with many tandem repeat, which is not seen in for instance CsVMV or BSV. It was assumed that TPV-like sequence had integrated repeatedly by illegitimate recombination into tobacco chromosomes until a copy number of about 1000 per diploid genome.

3.2 Material and methods

Potato tubers were bought from garden centres (see table 2.1) and grown in pots in a green house under standard conditions. Harvesting of leaves and roots for DNA isolation and chromosome squash preparations followed the protocols presented in chapter II: Materials and methods. PCR, cloning, sequencing, treatment of DNA for Southern hybridization, probe labelling and *in situ* hybridization were performed as in Chapter II. The Southern hybridizations were non-radioactive (55°C/2M Urea/0.5M NaCl). Sequences were aligned using ClustalW from EBI homepage (www.ebi.ac.uk), Neighbour Joining trees were generated from ClustalX and viewed in the programme TreeView. Dotplots for figure 3.9-3.11were created using pairwise sequence alignment in Megalign, Dotplots for figure 3.8 A and B were made with Dotter from Center for Genomics Research, Karolinska Institutet, Sweden (http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html; Sonnhammer and Durbin, 1995). GC content was determined using Oligocalc at www.justbio.com. Secondary fold structures were generated with mfold (www.bioinfo.rpi.edu/ applications/mfold/old/rna/form1.cgi) (Zucker *et al.*, 1999; Mathews *et al.*, 1999). In alignments, relevant parts of the database sequences TPV NTA238747 and TVCV AF190123 were used. A *gypsy*-like sequence PCR amplified from *Triticum aestivum* (TraeI) was used as outgroup.

3.3 Results

3.3.1 PCR, cloning and sequence analysis

Initially three primer pairs were designed from TPV-L (see table 2.2, of primer sequences and figure 2.1 for location) to amplify genomic DNA fragments from potato. Using standard PCR amplification conditions, two of the primer pairs gave products of approximately 800 bp and 1100 bp (figure 3.1). Lane 1 and 2 show amplification of 2µl and 5µl DNA respectively with the primer pair TPV3462-TPV4579, lane 5 and 6 show amplification of 2µl and 5µl DNA respectively with the primer pair TPV6241-TPV7041. The primer pair TPV890-TPV1909 did not show any amplification (lanes 3 and 4). There are reasons to believe that the forward TPV890 was functional as it was later used together with other primers and gave product, see below. The strong bands from lanes 1, 2, 5 and 6 were cut from the gel and the DNA was extracted. Figure 3.2 show these extracted fragments on a gel

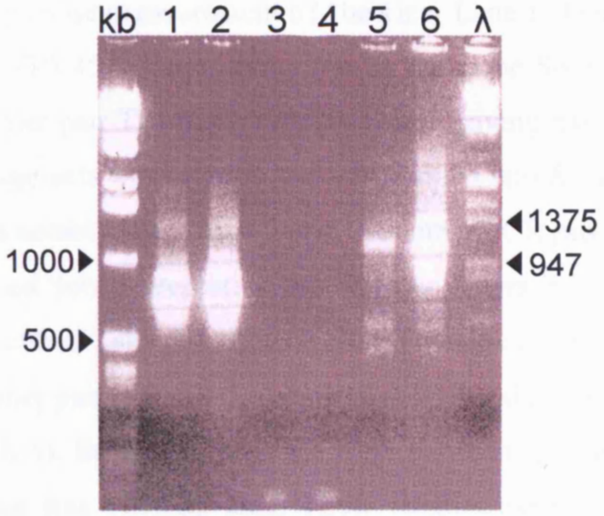


Figure 3.1. Gel image showing PCR amplification of potato genomic DNA with the primer pairs TPV6241-TPV7041 (lanes 1 and 2), TPV890-TPV1909 (lanes 3 and 4) and TPV3479-TPV4579 (lanes 5 and 6). In uneven numbered lanes 2 µl DNA was used in the PCR, in even numbered lanes 5 µl was used. Size markers are a 1kb ladder and lambda *Eco*RI/*Hin*dIII digest, with some sizes shown to the right and left of the image in bp.

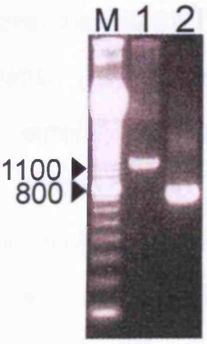


Figure 3.2. Gel image showing fragments cut from the gel in figure 3.1. Lane 1 from the primer pair TPV4362-TPV4579 giving rise to the clone SotuIII-1. Lane 2 from the primer pairs TPV6241-TPV7041 giving rise to SotuI clones. M is a 100 bp ladder, the size of the fragments are marked to the right.

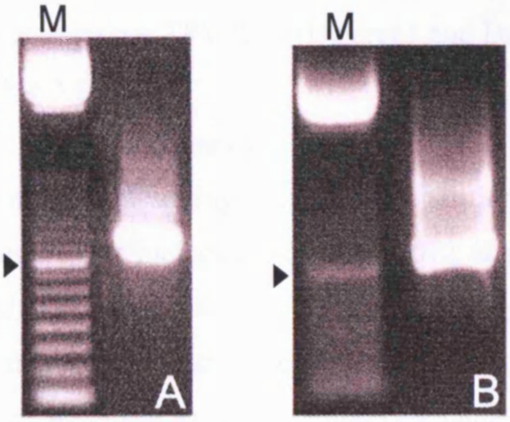


Figure 3.3. Gel images from the PCR amplification of potato genomic DNA with the primer pairs TPV6118-TPV7072 (A) and TPV6118-TPV7037 (B). Arrowheads to the left of images are marking 800 bp.

image giving a more precise measurement of the size. Lane 1, fragment from the primer pair TPV3462-TPV4579, later giving rise to the clone SotuIII-1 and lane 2, fragment from the primer pair TPV6241-TPV7041, later giving rise to SotuI clones. The extracted PCR fragments were cloned and transformed into *E. coli*, although the cloning efficiency was notably lower than found with other PCR products. All clones from potato were named Sotu representing the first two letters in genus and species name followed by a number. Table 3.1 lists the clones obtained: some 15 clones were obtained from one primer pair (SotuI-1-2, 4-5, 7-12, 14-17 and 21) and a single clone from the other (SotuIII-1). Sequences of SotuI were between 710 and 745 bp long, while the SotuIII clone was 1118 bp long. These lengths were approximately the same as for TPV, where the SotuI homologue is 800bp and the SotuIII homologue is 1117 bp. From another PCR reaction with the primers TPV6241-TPV7041 came the clone SotuI-22 which was shorter than the other (667 bp), and not homologous to TPV but 71% homologous to part of the intergenic region of the *gypsy* retrotransposon *Bagy-1*.

In a second series of experiments, four more primers were designed in the same region as TPV6241-TPV7041, for the possibility of nesting primer pair sets (table 2.2, figure 2.1). Three of the primers were used for semi-nested PCR (reamplifying a small aliquot of the first round product using one original primer and one internal to a primer used in the first round, to improve sensitivity and specificity). This amplification, with TPV6118-TPV7072 (first round) and TPV6118-TPV7037 (second round) gave bands of around 900 bp (figure 3.3); suitable products were sequenced directly or cloned. The clone from a TPV6118-TPV7072 amplification, Sotu6118/7072, was 870 bp long, the expected size from TPV being 954bp. PCR products from semi-nested PCR, sequenced directly with the reverse primer TPV7037 (Desiree1 and Desiree2), were 470 and 443 bp respectively (table 3.1).

The sequences of SotuI and Sotu6118/7072 clones were aligned in ClustalW using default settings (figure 3.4), the left hand primer and most of the right hand primer were included. Sotu6118/7072 as the only one had primers further upstream and downstream so this clone is not very homologous to the others in the shown primer regions. Part of sequences from TPV and TVCV were included, and the clone TraeI was used as an out-group (see table 4.1). A phylogram was generated with the Neighbour Joining method from the alignment and showed that the SotuI clones

Table 3.1. Table of clones obtained from potato (Sotu) and two PCR products, Desiree-1 and Desiree-2, and the length of the sequences. H_2O control, whether a negative control was included in the PCR. The forward and reverse primer used in the PCR. PCR condition gives the annealing temperature.

Clone/PCR	Fragment	Sequence length	H_2O control	F-primer	R-primer	PCR condition
Sotu6118/7072	5' -3'	870bp	yes, neg.	TPV6118	TPV7072	50°C
SotuI-1	5' -3'	733 bp	no	TPV6241	TPV7041	56°C
SotuI-2	5' -3'	741 bp	no	TPV6241	TPV7041	56°C
SotuI-4	5' -3'	735 bp	no	TPV6241	TPV7041	56°C
SotuI-5	5' -3'	740 bp	no	TPV6241	TPV7041	56°C
SotuI-7	5' -3'	741 bp	no	TPV6241	TPV7041	56°C
SotuI-8	5' -3'	733 bp	no	TPV6241	TPV7041	56°C
SotuI-9	5' -3'	734 bp	no	TPV6241	TPV7041	56°C
SotuI-10	5' -3'	741 bp	no	TPV6241	TPV7041	56°C
SotuI-11	5' -3'	734 bp	no	TPV6241	TPV7041	56°C
SotuI-12	5' -3'	751 bp	no	TPV6241	TPV7041	56°C
SotuI-14	5' -3'	733 bp	no	TPV6241	TPV7041	56°C
SotuI-15	5' -3'	734 bp	no	TPV6241	TPV7041	56°C
SotuI-16	5' -3'	742 bp	no	TPV6241	TPV7041	56°C
SotuI-17	5' -3'	768 bp	no	TPV6241	TPV7041	56°C
SotuI-21	5' -3'	733 bp	no	TPV6241	TPV7041	56°C
SotuI-22	5' -3'	667 bp	yes, neg.	TPV6241	TPV7041	56°C
SotuIII-1	5' -3'	1118 bp	no	TPV3462	TPV4579	56°C
Desiree-1	3'	416 bp	yes, neg.	TPV6118	TPV7017	50-52°C
Desiree-2	3'	419 bp	yes, neg.	TPV6118	TPV7017	50-52°C

Figure 3.4. Alignment of whole potato clones (Sotu) inclusive the relevant parts of TPV (Tobacco pararetrovirus-like sequence, NTA238747) and TVCV (*Tobacco vein clearing virus*, AF190123). The sequence TraeI with high similarity to a *gypsy* type retrotransposon was used as out-group. Primers are marked at each end and the nucleotide difference between the two potato groups are marked with grey.

```
                Primer TPV6241
SotuI-9         ATGCAACCAACTACCAGAGCAATACGGAAGAACTTTATTTCACCAGAGATAATGACAAATTATTGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-12        ATGCAACCAACTACCAGAGCAWTACGGAAGAACTTTATTTCACCAGAGATAATGACAAATTATTGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-14        ATGCAACCAACTACCAGAGCAATACGGAAGAACTTTATTTCACCAGAGATAATGACAAATTATTGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-11        ATGCAACCAACTACCAGAGCAATACGGAAGAACTTTATTTCACCAGAGATAATGACAAATTATTGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-21        ATGCAACCAACTACCAGAGCAATACGGAAGAACTTTATTTCACCAGAGATAATGACAAATTATTGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-8         ATGCAACCAACTACCAGAGCAATACGGAAGAACTTTATTTCACCAGAGATAATGACAAATTATTGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-17        ATATGGTCGACCTGCAGGCGGACGCGGTAGAGCTTTATYCCCCCAAAGATAATGACAAATTATGGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-1         ATGCAACCAACTACCAGAGCAATACGGAAGAACTTTATTTCACCAGAGATAATGACAAATTATTGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-4         ATGCAACCAACTACCAGAGCAATACGGAAGAACTTTATTTCACCAGAGATAATGACAAATTATTGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-15        ATGCAACCAACTACCAGAGCAATACGGAAGAACTTTATTTCACCAGAGATAATGACAAATTATTGCAAAACTATTGGACACAGATACCCAGATCATCAATGCTCAAAATGCTATGGAGAA 120
SotuI-2         ATGCAACCAACTACCAGAGCAATACGAAAGAATTTTATTTCACCAGAGATAATGACAAAATTTTGCAAGACTATTGGACACAGATACCCAGATCACCAATGTTCAGAATGCTATGGAGAA 120
SotuI-10        ATGCAACCAACTACCAGAGCAATACGAAAGAATTTTATTTCACCAGAGATAATGACAAAATTTTGCAAGACTATTGGACACAGATACCCAGATCACCAATGTTCAGAATGCTATGGAGAA 120
SotuI-5         ATGCAACCAACTACCAGAGCAATACGGAAGAATTTTATTTCACCAGAGATAATGACAAAATTTTGCAAGACTATTGGACACAGATACCCAGATCACCAATGTTCAGAATGCTATGGAGAA 120
SotuI-7         ATGCAACCAACTACCAGAGCAATACGAAAGAATTTTATTTCACCAGAGATAATGACAAAATTTTGCAAGACTATTGGACACAGATACCCAGATCACCAATGTTCAGAATGCTATGGAGAA 120
SotuI-16        ATCCAACCAACTACCAGAGCAATACGAAAGAATTTTATTTCACCAGAGATAATGACAAAATTTTGCAAGACTATTGGACACAGATACCCAGATCACCAATGTTCAGAATGCTATGGAGAA 120
TPV             ATGCAACCAACTACCAGAGCATTGAAAAAAGAATTTATTTCAAACGAGTTATTAACAAGATACTGCAAACTAGTAGGACACAAATATCCAGACCACATATGTTCAAAGTGCAACGGAGAT 120
TVCV            CAACCTACAACCACAAGAGCTATAAGGAGGAATTTCATCTCCCCAGATCTGTTAACAAGATACTGCAAGCTAATTAGTCACAAATATCCAGACCACCTATGCTCAAAATGCAAAGGAGAA 120
Sotu6118        GAACGACCTACAACACGAGCACTAAAGATAGGATTTATTTCAGCAGAGTTATTAACTAGGTACTGCAAGTTAATAGGACATAAGTACCCAGATCATATATGTTCAAAATGTAATCGCGAA 120
TraeI           ATGCAACCAACTACCAGAGC--------------------------------------------------------------------------ATACAAGCAACARGCACATGCGGAAG 46
                       * **     * *        *     ** **  *     *  *  * ** *  *  *****    * *   ** *  ** ***** **   *** *** * **  *  * **

SotuI-9         GACAATGTAATTACAGCTGTCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAAAGCAAGA--------TAAAATAAAGTA---------AAGTTGT-----------TTGTCTT 207
SotuI-12        GACAATGTAATTACAGCTGTCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAAAGCMAGA--------TAAAATAAAGTA---------AAGTTGT-----------TTGTCTT 207
SotuI-14        GACAATGTAATTACAGCTGTCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAAAGCAAGA--------TAAAATAAAGTA---------AAGTTGT-----------TTGTCTT 207
SotuI-11        GACAATGTAATTACAGCTGTCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAAAGCAAGA--------TAAAATAAAGTA---------AAGTTGT-----------TT-TCTT 206
SotuI-21        GACAATGTAATTACAGCTGTCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAAAGCAAGA--------TAAAATAAAGTA---------AAGTTGT-----------TTGTCTT 207
SotuI-8         GACAATGTAATTACAGCTGTCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAA-GCAAGA--------TAAAATAAAGTA---------AAGTTGT-----------TTGTCTT 206
SotuI-17        GACAATGTAATTACTGCTGTCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAAAGCAAGA--------TAAARTAAAGTG---------AAGTTGT-----------TTGTCTT 207
SotuI-1         GACAATGTAATTACAGCTGTCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAAAGCAAGA--------TAAAATAAAGTA---------AAGTTGT-----------TTGTCTT 207
SotuI-4         GACAATGTAATTACAGCTGCCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAAAGCAAGA--------TAAAATAAAGTA---------AAGTTGT-----------TTGTCTT 207
SotuI-15        GACAATGTAATTACAGCTGCCCAGCTA-GAATAAA--AAGC--CAACGACAAGATAAAAAGCAAGA--------TAAAATAAAGTA---------AAGTTGT-----------TTGTCTT 207
SotuI-2         GATAACGTGATTCCTGATGTCCAGCTA-GAGTAAA--AGCCAACAACGACAAGATAAGAAGCAAGAA-------TGACACCCGACA---------AGAGAA------------------- 202
SotuI-10        GATAACGTGATTCCTGATGTCCAGCTA-GAGTAAA--AGCCAACAACGACAAGATAAGAMGCAAGAA-------TGACACCCGACA---------AGAGAA------------------- 202
SotuI-5         GATAACGTGATTCCTGATGTCCAGCTA-GAGTAAA--AGCCAACAACGACAAGATAAGAAGCAAGAA-------TGACACCCGACA---------AGAGAA------------------- 202
SotuI-7         GATAACGTGATTCCTGATGTCCAGCTA-GAGTAAA--AGCCAACAACGACAAGATAAGAAGCAAGAA-------TGACACCCGACA---------AGAGAA------------------- 202
SotuI-16        GATAACGTGATTCCTGATGTCCAGCTA-GAGTAAA--AGCCAACAACGACAAGATAAGAAGCAAGAA-------TGACACCCGACA---------AGAGAA------------------- 202
TPV             GATAACTATGTACCAGAAGTCCAACTA-GAATGA---AGGTATCAACAAGAAGACAAAAGAAGAAGACAGCTACTGGAAACATATAATGTAAAGTAAATAGTACTAGTCACATTCATGAA 236
TVCV            GATAACATTGTTCCAGACGTACAACTG-GAATAATTCAGATAG-AAGAAAACGACAAAACAACAAGATA--TGACAGAAAGATCTACTTT----TATAGATTCCTTTTT---CTTTTGTA 229
Sotu6118        GATAATATAATTCCAGATGTCCAACTA-GAATGAT--CAACAGGTGCAAAGAGACAAGAAGATAAA------GACAAGAAGATA-----------AGAGAG--------------ACACG 206
TraeI           CTTAACTTGTCTGAGTACAGACAACTACAAATGAAGAAGGC-TCAGAAGCCTGATTACCTACAAGACCC--TCCCAAGGTACAAGATCGTAGCTAAGGTAAC------------------ 145
                ** **     *  * *  *  ** **  ** * *              *   ** ** *    *                               *

SotuI-9         ATTATTTG-------ACTTATTATAAAAGATA-GTGGGATTAGGAATCTTAT-GTAAATTAGTGGACAAG--TGTMAAGTAGT-GTGCATTATTAAAACGTAGTGCTC--TGTAAAG--- 310
SotuI-12        ATTATTTG-------ACTTATTATAAAAGATA-GTGGGAT-AGGAATCTTAT-GTAAATTAGTGGACAAG--TGTAAAGTAGT-GTGCATTATTAAAACGTACTGCTC--TGTAAAGAAC 312
SotuI-14        ATTATTTG-------ACTTATTATAAAAGATA-GTGGGATTAGGAATCTTAT-GTAAATTAGTGGACAAG--TGTAAAGTAGT-GTGCATTATTAAAACGTAGTGCTC--TGTAAAG--- 310
SotuI-11        ATTATTTG-------ACTTATTATAAAAGATA-GTGGGATTAGGAATCTTAT-GTAAATTAGTGGACAAG--TGTAAAGTAGT-GTGCATTATTAAAACGTAGTGCTC--TGTAACA--- 309
SotuI-21        ATTATTTG-------ACTTATTATAAAAGATA-GTGGGATTAGGAATCTTAT-GTAAATTAGTG-ACAAG--TGTAAAGTAGT-GTGCGTTATTAAAACGTAGTGCTC--TGTAA-A--- 308
SotuI-8         ATTATTTG-------ACTTATTATAAAAGATA-GTGGGATTAGGAATCTTAT-GTAAATTAGTGGACAAG--TGTAAAGTAGT-GTGCATTATTAAAACGTAGTGCTC--TGTAA-A--- 308
SotuI-17        ATTATTTG-------ACTTATTATAAAAGATA-GTGGGATTAGGAATCTTAT-GTAAATTAGTGGACAAG--TGTAAAGTAGT-GTGCATTATTAAAACGTAGTGCTC--TGTAAAG--- 310
SotuI-1         ATTATTTG-------ACTTATTATAAAAGATA-GTGGGATT-GGAATCTTAT-GTAAATTAGTGGACAAG--TGTAAAGTAGT-GTGCATTATTAAAACGTAGTGCTC--TGTAAAG--- 309
SotuI-4         ATTATTTG-------ACTTATTATAAAAGATA-GTGGGATTAGGAATCTTAT-GTAAATTAGTGGACAAG--TGTAAAGTAGT-GTGCATTATTAAAACGTAGTGCTC--TGTAAAG--- 310
SotuI-15        ATTATTTG-------ACTTATTATAAAAGATA-GTGGGATTAGGAATCTTAT-GTAAATTAGTGGACAAG--TGTAAAGTAGT-GTGCATTATTAAAACGTAGTGCTC--TGTAAAG--- 310
SotuI-2         AAAAAAAA-------GCCGAAT-CAAAAGAT--GTAAAGTTGTTTGTCTTATTATAAAATACT--ATAAT--GGTAAATTATCAGTAGACAAGTGTAAAGGCGCATTAG-TGTAAAG--- 304
SotuI-10        AAAAAAAH-------GCGGAAT-CAAARGAT--GTMAAGTTGTTTGTCTTATTATAAAATAGT--ATAAT--TGTAAATTATCAGTAGACAAGTGTAAAGGCGCATTAG-TGTAAAG--- 304
SotuI-5         AAAAAA---------GCAGAAT-CAAAAGAT--GTAAAGTTGTTTGTCTTATTATAAAATAGT--ATAAT--TGTAAATTATCAGTAGACAAGTGTAAAGGCGCATTAG-TGTAAAG--- 302
SotuI-7         AAAAAAAA-------GCAGAAT-CAAAAGAT--GTAAAGTTGTTTGTCTTATTATAAAATAGT--ATAAT--TGTAAATTATCAGTAGACAAGTGTAAAGGCSCATTAG-TGTAAAG--- 304
SotuI-16        AAAAAAAA-------GCAGAAT-CAAAAGAT--GTAAAGTGGTTGGTCTTATTATAAAATAGT--ATAAT--TGTAAATTATCAGTAGACAAGTGTAAAGGCGCATTAG-TGGGAAG--- 304
TPV             CAGTAAAGGTCGTTCATGAATAGTAAGAGTCATGTAAAAAAGTCGTAAAGTAAATAGTATATGTCATAATCATGAACAGTAAAGGTCGTTCATGAATAGTAAGAGTCAT--GTAAA---- 350
TVCV            AAGAAATGTTATTTTGTTTTTTATAAAAGTCGTAAAGAATAGT--TTACTTTATTAAAGTAAGAGTCATTCATGAACAGTAAAGGTCGTTCATGAATAGTAAGAGTCGTTTGTAAATAG- 346
Sotu6118        AAGATAAG-------------AAGAAAAGATGTAAAAAAAAAA----AAGATAGTGACATAAA----------GTAAAGTAAATGCTTTT---------------TCTTTTCTGAA---- 280
TraeI           AAGCTAAACGTCAAAGTCCACACGGAACTACTAGAGAGACTGACSTCTCTAT-GCAAAACATACAATAAG---CAAACGTGAGTGCAAAT-----GTACCTAGCACGACTTACATCA--- 253
                                        ** **                         *    **            *   * **   *
```

Figure 3.4. *Continued.*

```
SotuI-9     ----------------TAATAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTAGAAATAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 402
SotuI-12    GTATTGGCTGTAAAGGTATAAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTAGAAATAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 420
SotuI-14    ----------------TAATAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTASAAMTAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 402
SotuI-11    ---------------GTAATAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTAGAAATAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 402
SotuI-21    ---------------GTAATAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTAGAAATAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 401
SotuI-8     ---------------GTAATAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTAGAAATAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 401
SotuI-17    ----------------TAATAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTAGAAATAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 402
SotuI-1     ----------------TAATAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTAGAAATAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 401
SotuI-4     ----------------TAATAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTAGAAATAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 402
SotuI-15    ----------------TAATAATTAGGAGACAA----ATGTAAAGTAGGAATAGTAT----AGGACAGATGTAAAGTAGAAATAGTATA--GCATCTCTACTTGTATAAATAG--AGAGC 402
SotuI-2     ----------------CAATAAT-AGAGGACAG----ATGTAGAGTAAGAATAGTAT----AGGACAGATGTAAAGGAGAAAT--------CTACATTCTCCTCTATAAATAG--GAAGC 389
SotuI-10    ----------------CAATAAT-AGAGGACAG----ATGTAAAGTAAGAATAGTAT----AGGACAGATGTAAAGGAGAAGT--------CTACATTCTCCTCTATAAATAG--GAAGC 389
SotuI-5     ----------------CAATAAT-AGAGGACAG----ATGTAAAGTAAGAATAGTAT----AGGACAGATGTAAAGGAGAAAT--------CTACATTCTCCTCTATAAATAG--GAAGC 387
SotuI-7     ----------------CAATAAT-AGAGGACAG----ATGTAAAGTAAGAATAGTMT----AGGACAGATGTAAAGGAGAAAT--------TTACATTCTCCTCTATAAATAG--GAAGC 389
SotuI-16    ----------------CCATTAT-AGAGGACAG----ATGTTAAGTAAGGATAGTATT---AGGACAGATGTTAAGGAGGAAT--------CTACATTCTCCTCTATAAATAG--GAAGC 390
TPV         ----AAAGTCGTAAAGTAAATAGTACGAGTCATAATCATGAACAGTAAAGGTCGTTCATGAATAGTAGGAGTCATTTGTAAACAGTAAGAGTCG-TTTTAATTTTCTTTATA-----TAG 460
TVCV        --TAAGAGTCGTTT-GTAAATAGTAAGAGTCGT--TTGTAAATAGTAAGAGTCGTT-------AGTAAGAGTCGTTTGTGAATAGTAAGAGTCAATTTTTGTATTATAAATAGCAGTTCA 454
Sotu6118    -----AGGCCAAAGTGTGAATAGTAAAGGCCA----ATGAATAGTAGAAG--------------GCAAG-------TGTAAACAGTAGGAGTCATTTATTTTCCTATATATAG--GCATA 368
TraeI       ----------------GAACTATCTACATATGC----ATCAGTATCAACAAAGGGGT------GGTGGAGTTTAACTGCARCAAG------CTAGCTTTGACTCGGTGGCTAACCTGAAC 341
                                     *  *   * *       *    ****                   *             *               *       * *  ***
```

```
SotuI-9     CATTAGGCACATCTAAGGCAAGACTTTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCAATAAAATATC----AAGTTTCCAATCTAAGCTATGGATCAAAA------- 508
SotuI-12    CATTAGGCACATCTAAGGCAAGACTTTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCAATAAAATATC----AAGTTTCCAATCTAAGCTATGGATCAAAA------- 526
SotuI-14    CATTAGGCACATCTAAGGCAAGACTTTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCAATAAAATATC----AAGTTTCCAATCTAAGCTATGGATCAAAA------- 508
SotuI-11    CATTAGGCACATCTAAGGCAAGACTTTCTGGACGGAAWGCAAGCCTCC-TTGTACACCCCAATA--CTCAATAAAATATC----AAGTTTCCAATCTAAGCTATGGATCAAAA------- 508
SotuI-21    CATCAGGCACATCTAAGGCAAGACTTTCTSGACGGWAAGCAAGCCTCC-TTGTAAACCCCAATA--CTCAATAAAATATC----AAGTTTCCAATCTAAGCTATGGATCAAAA------- 507
SotuI-8     CATTAGGCACATCTAAGGCAAGACTTTCTCGACGGAAAGCAAGCCTCC-TTGTAAACMRRAATA--CTCAATAAAATATC----AAGTTTCCAATCTAAGCTATGGATCRAAA------- 507
SotuI-17    CATTAGGCACATCTAAGGCAAGACTTTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCAATAAGATATC----AAGTTTCCAATCTAAGCTATGGATCAAAA------- 508
SotuI-1     CATTAGGCACATCTAAGGCAAGACTTTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCAATAAAATATC----AAGTTTCCAATCTAAGCTATGGATCAAAA------- 507
SotuI-4     CATTAGGCACATCTAAGGCAAGACTTTCTCGACGGAAAGCAAGCCTCGGTTGTAAACACCTATA--CTCAATAAAATATG----GAGTTTCCAATCTAAGCTATGGATCAGAA------- 509
SotuI-15    CATTAGGCACATCTAAGGCAAGACTTTCTCGACGGAAAGCAAGCCTCC-TTGTAAACCCCAATA--CTCAATAAAATATC----AAGTTTCCAATCTAAGCTATGGATCAAAA------- 508
SotuI-2     CATTTAGGTAATCTAAGGCAAGACTTTCCCGACGGAAAGCAAGCCTCT-TTGTAAACAAAAATATTCTCAATAAAATATCA---AAGTTTTCAAGCTAAGTTATGAATCAAAATACAGTT 505
SotuI-10    CATTTAGGTAATCTAAGGCAAGACTTTCCCGACGGAAAGCAAGCCTCT-TTGTAAACAAAAATATTCTCAATAAAATATCA---AAGTTTTCAAGCTAAGTTATGAATCAAAATACAGTT 505
SotuI-5     CATTTAGGTAATCTAAGGCAAGACTTTCCCGACGGAAAGCAAGCCTCT-TTGTAAACAAAAATATTCTCAATAAAATATCA---AAGTTTTCAAGCTAAGTTATGAATCAAAATACAGTT 503
SotuI-7     CATTTAGGTAATCTAAGGCAAGACTTTCCCGACGGAAAGCAAGCCTCT-TTGTAAACAAAAATATTCTCAATAAAATATCA---AAGTTTTCAAGCTAAGTTATGAATCAAAATACAGTT 505
SotuI-16    CATTTAGGTAATCTAAGGCAAGACTTTCCCGACGGAAAGCMAGCCTCT-TTGTAAACCAAAATATTCTCCATAAAATWTCC---AAGTTTTCCAGGTTAGTTWTGGATCCAAAWACCGTT 506
TPV         AATGTAAA---TCTGAGGAAGGAGGACATCCTC---AGACATCC--TCATCCTCACCTTCTCTC--TCTTATCTTCTCTCAATAAAATATCTGATTATATACAAACTTCTGAAAGCTATG 570
TVCV        AATGTGAA---TAAAAAACAGGCTGCAGTTTTC---AAGCATCCAACAATTCCTCTCTCTTCTC--TCTAATATATTTT-----GCAGATATAAGCATACGAAAGCTATGGAACAACAAG 561
Sotu6118    GATGTAAACATTGTAAAGCAAGCCTTCACTACG---AAGCAAGC--TCATTTGTAAAAACACTC--TTAAATATATTTGC-----ATTTTAGAAACCTAATGGAGCAACTAGATAAATCA 476
TraeI       TACG------ATGCTATGTAACTCTTTTGTGGTGRTGCACACGAGTCC-----ACATATTCGCCATATCAATACAMANAT----ATGAATCCGCTSCCGTCTCCATACGAAAA------- 439
             **        *   *   * *              *  *   *    *             *      **    *             *   *    *             *
```

```
SotuI-9     -----------TATTCCAGATATCACACATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACA 617
SotuI-12    -----------TATTCCAGATATCACACATGTATATCTATTTATGATTATGCATAG-TAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACA 634
SotuI-14    -----------TATTCCAGATATCACACATGTATATCTATTTATGATTATGCATAGCTAA-TCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACA 616
SotuI-11    -----------TATTCCAGATATCACACATGTATATCTATTTATGGTTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACA 617
SotuI-21    -----------TATTCCAGATATCACACATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACA 616
SotuI-8     -----------TATTCCAGATATCACACATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACA 616
SotuI-17    -----------TATTCCAGATATCACACATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACA 617
SotuI-1     -----------TATTCCAGATATCACACATGAATATCTATTAATGATTATGC-TATTTTTATCTGTTAATTCTTGAATATCATCAGCATGATTAAATAAGTTTATATTAGTTATTATACA 615
SotuI-4     -----------TATTCCAGATATCACACATGTATATCTATTTATGATTATGCATAGCTCAATCTGTTAATTCTTGAATATCATTAGCATGCTTAAATAAGTTTATACTAGTTATTATATA 618
SotuI-15    -----------TATTCCAGATATCACACATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGCTTAAATAAGTTTATACTAGTTATTATATA 617
SotuI-2     ATGGATTAAAATACACCAGATATCGAACATGTATATCTGTCTATGATTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATA 625
SotuI-10    ATGGATTAAAATACACCAGATATCGAACATGTATATCTGTCTATGATTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATA 625
SotuI-5     ATGGATTAAAATACACCAGATATCGAACATGTATATCTGTCTATGATTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATA 623
SotuI-7     ATGGATTAAAATACACCAGATATCGMACATGTATATCTGTCTATGATTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATA 625
SotuI-16    ATGGAWTTAAATACCCCCGAWWTCGGACMAGGAWATCYGTTTAWGATTAAGCATAGCTWAATTTATTGATTCCTTAATATCATTAGCATGATTGGATAAGTTTATATTAGTTATTATAWA 626
TPV         GAGCATTCAAACGAGTTACAAA---ACCAGGTAAGTTTATTTATAATCATGTTTAACTAAACTCTTATATTCCTTATAATCTCT-TGAATATCATAATTGTTTAT------CATGTTATA 680
TVCV        GAG-ATAGCAGAAATCTACAAA---AACAGGTATATCATTTTCCAACTATGAACAGCTAAATTACTATATTCCTGATAATTTCTATGTAAATTATAATTGTTCAT------CATACTATA 671
Sotu6118    GATATCACGGAGAAGCAAGAGG---CAAAGGTATATTGTTTATTAAATATGCAGAACTAAATTCATATATTCCTGAATATCATAAATATGATTGAATTAGTTTATGCTAGTTATTATGTA 593
TraeI       -----------ARCCATCCATAGCACTCACGCTTATCT-TGCGTACTTTAGAGTATCCACTWCCACATGTCTATGAACTATGCAAAGG-GTCCGAGTTTCCATATMCGAGGAATCCGGCT 546
                              *          * *  *   *        * *   *  *       *  **** * *  **            *   *   *** **       **  *  *
```

Primer TPV7037

```
SotuI-9     CTAGAATTATT-TGAAGTATGTTTTCTA-GTTTGCTAACATGAGGAAAAAAAGAAAAGCTT----CCAAAACTATGTCATCCTAAAGT-TGAAACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 734
SotuI-12    CTAGAATTATT-TGAAGTATGTTTTCTA-GTTTGCTAACATGAGGAAAAAAAGAAAAGCTT----CCAAAACTATGTCATCCTAAAGT-TGAAACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 751
SotuI-14    CTAGAATTATT-TGAAGTATGTTTTCTA-GTTTGCTAACATGAGGAAAAAAAGAAAAACTT----CCAAAACTATGTCATCCTAAAGT-TGAAACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 733
SotuI-11    CTAGAATTATT-TGAAGTATGTTTTCTA-GTTTGCTAACATGAGGACAAAAAGAAAGACTT----CCAAAACTATGTCATCCTAAAGT-TGAAACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 734
SotuI-21    CTAGAATTATT-TGAAGTATGTTTTCTA-GTTTGCTWACATGAGGAAAAAAAGAAAGACTT----CCAAAACTATGTCATCCTAAAGT-TGAAACCA-TCGGC--AAGGAAGCCGTT-TAGGGGAGTW 733
SotuI-8     CTAGAATTATT-TGAAGTATGTTTTCTA-GTTTGCTAACATGAGGAAAAAAAGAAARACTT----CCAAAACTATGTCATCCTAAAGT-TGAAACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 733
SotuI-17    CTAGAATTATT-TGAAGTATGTTTTCTA-GTTTGCTAACATGAGGAAAAAAAGAAAAACTT----CCAAAACTATGTCATCCTAAAGT-TGAAACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 734
SotuI-1     CTAGAATTATT-TGAAGTATGTTTTGTA-GTTTGCTTACATGAGGAAGAAAAGATATACTKT---CCAAAAATAGTTCATCCTTAAGT-TGAAACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 733
SotuI-4     CTAGAATTATT-TAAAGCATGTTTTCTA-GTTTGTTWATATGAGGAAAAAAGGARAGATTT----CCAAAACTATTCCATCCTAAAGT-TGAWACGA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 735
SotuI-15    CTAGAATTATT-TAAAGCATGTTTTCTA-GTTTGTTAATATGAGGAAAAAAGGAAAAATTT----CCAAAACTATGCCATCCTAAAGT-TGACACGA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 734
SotuI-2     CTAGAATTATT-CAAAGCATGTTT-ATA-GTTTGCTAACATGAGGAAAAAAGGAAAAACTT----CCAAAACTATGCCATCCTAAGGT-TGAAACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 741
SotuI-10    CTAGAATTATT-CAAAGCATGTTT-ATA-GTTTGCTAACATGAGGAAAAAAGGAAAAACTT----CCAAAACTATGCCATCCTAAGGT-TGAAACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 741
SotuI-5     CTAGAATTATT-CAAAGCATGTTT-ATA-GTTTGCTAACATGAGGAAAAAAGGAAAAACTT----CCAAAACTATGCCTTCMWAAGGTGTGAAACCA-TAGGC--AAGGAAGCGGTT-TAGGGGAGTA 740
SotuI-7     CTAGAATTATT-CAAAGCATGTTT-ATA-GTTTGCTAACATGAGGWMAAAAGGAAAGACTT----CCAAAACTATGCCATCCTAAGGT-TGATACCA-TAGGC--AAGGAAGCCGTT-TAGGGGAGTA 741
SotuI-16    CTAGATTAATT-CAAAGCATGTTT-ATA-GTTTGCTAACATGAGGAAAAAAGGAAAAACTT----CCBAGACTATGCCATCCTAAGGT-TGAWACCA-TMGGC--AAGGMATCCGTT-TAGGGGAGTA 742
TPV         GTACTGTTATA-AAGAACATCTTGAGAAAGTTTGCCTGTCTAATAATGCTGAGAGTAGTTAAG--CCCAGACTATGCCATCCTAAAGT-GGAAACCAGTAGGC--AAGGAAGCCGTT-TAGGGGAGTA 801
TVCV        GTACTGTTATA-AAAATCATCTTGAGAAAGTTTGCCATGATAATA-TGCTAATAGTAGGTAAA--CCCAGACTATGCCATCCTAATGT-TGAAACCGGTAGGC--AGAGAAGCCGTT-TAGGGGAGTA 791
Sotu6118    TTAGAATTATACGAAATCATGCTTTCTA-GTTTATTAGCA-AACAGGAAAAGCGAGAAACCTT--CCATAACTATGCCATCCTAAATT-TGAAACCAGAAAGGCAAGAGAAGCCGTGATAGGGGAGTA 716
TraeI       ATTCTAATAGATAATGATAACCCTGGCAGGGGTGACTTCTTCACACACGCTCTCGCCACTTATCGCCCTATWCACGTCATGT--ACCTCGGCAACCTTCAAGC----GGAAGCCGTT-TAGGGGAGTA 667
             **   * **     *  **  *    * ****         *                      **   * **   * **   *  *  ** **      *      * * ****  **********
```

Most of TPV7041

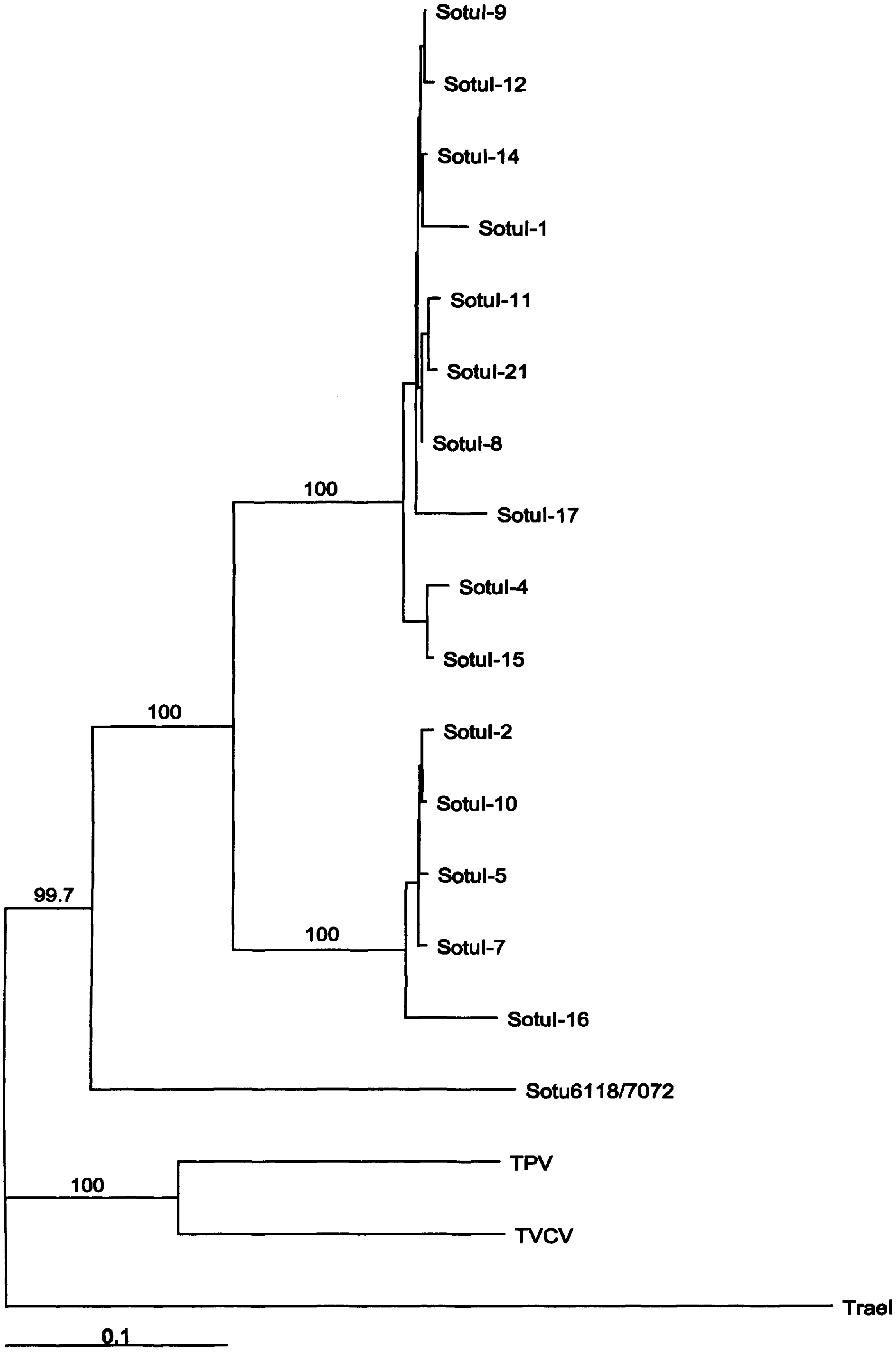


Figure 3.5. Phylogram from alignment of potato clones (Sotu) as seen in figure 3.3. The origin of clone names are in table 3.1. Numbers on major branches are the bootstrap values in percentages based on 1000 trials. Bar indicates the branch length corresponding to a 10% difference.

Table 3.2. Similarities in percentages between individual clones from the alignment (figure 3.4). The different colours correspond to the groups on the phylogram (figure 3.5).

	Sotul-9	Sotul-12	Sotul-14	Sotul-1	Sotul-11	Sotul-21	Sotul-8	Sotul-17	Sotul-4	Sotul-15	Sotul-2	Sotul-10	Sotul-5	Sotul-7	Sotul-16	Sotu6118	TPV	TVCV
Sotul-9																		
Sotul-12	98																	
Sotul-14	99	98																
Sotul-1	97	96	96															
Sotul-11	98	97	98	96														
Sotul-21	98	97	98	96	98													
Sotul-8	99	98	98	96	98	98												
Sotul-17	94	94	94	92	93	93	94											
Sotul-4	96	95	96	95	96	96	96	91										
Sotul-15	97	96	97	95	97	97	97	93	98									
Sotul-2	82	81	82	80	81	81	82	78	80	82								
Sotul-10	82	81	82	80	81	81	82	78	80	82	98							
Sotul-5	82	81	82	80	81	81	82	78	80	82	98	98						
Sotul-7	82	81	82	80	81	81	82	78	81	82	98	98	98					
Sotul-16	77	76	77	75	76	76	77	73	76	77	92	92	92	92				
Sotu6118	66	66	66	64	64	64	65	63	65	66	69	69	68	68	66			
TPV	66	65	66	65	66	65	66	61	65	66	65	65	65	64	62	67		
TVCV	64	63	64	63	64	64	64	61	63	64	62	63	63	62	61	66	72	

clustered in two groups (figure 3.5). The two groups differed by 184 base substitutions and indels (shown as grey marking between the primers, figure 3.4). Within the groups the similarity was high - between 91 and 99% (table 3.2). Overall the similarity between SotuI clones ranged from 80% to 99%. Sotu6118/7072 was separated on another branch being 63-69% identical to the SotuI clones. In the bottom TPV and TVCV are on the same major branch, 72% identical to each other.

Another alignment was generated which included the reverse sequenced PCR products. Removal of the priming sites and shortening of the sequences was necessary to make the alignment fit with the shortest sequence (figure 3.6). The alignment again included TPV, TVCV and TraeI. The SotuI clones clustered in the same two groups, with no notable differences (figure 3.7). The trimming of the sequences from 700+ bp to barely 400 bp did not affect the major phylogenetic grouping, which shows there was a stable relatedness between the sequences. In the trimmed alignment the sotu6118/7072 clone clustered together with the PCR products Desiree-1 and -2 with about 96% similarity to these (table 3.3). These three latter sequences were amplified with the same primers and differ from the SotuI clones with about 130 base substitutions and indels (shown as grey in figure 3.6). The overall similarity between potato PRV-L sequences ranged from 63% to 100% (table 3.3). For both phylograms (figure 3.5 and 3.7) the bootstrap values on the major branches shows that the trees are well supported.

An attempt was made to PCR amplify longer fragments by combining primer pairs and using conditions for amplification of longer sequences (Boehringer, Expand long template PCR system). The product was shot gun cloned and several large inserts were obtained. One was 7-8 kb which could be a nearly complete virus-like sequence compared to TPV. A part of it was sequenced and found to be homologous to known pararetroviruses but further sequencing was halted because of the very repetitive nature of the sequence. Other attempts to produce longer fragments gave weak PCR products of the expected size, but the concentration was too low for cloning or sequencing.

3.3.2 Homologies to database sequences

Searching for homologous sequences in the EMBL nucleotide database with Fasta using potato sequences as query always found high homology to TPV or TVCV. SotuIII, having homology to the very conserved region of RT, also had a high

```
SotuI-14    CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-SAAMTAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATTAGGCACATCTAAG-GCAAGAC----T 103
SotuI-17    CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-GAAATAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATTAGGCACATCTAAG-GCAAGAC----T 103
SotuI-9     CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-GAAATAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATTAGGCACATCTAAG-GCAAGAC----T 103
SotuI-12    CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-GAAATAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATTAGGCACATCTAAG-GCAAGAC----T 103
SotuI-11    CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-GAAATAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATTAGGCACATCTAAG-GCAAGAC----T 103
SotuI-21    CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-GAAATAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATCAGGCACATCTAAG-GCAAGAC----T 103
SotuI-8     CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-GAAATAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATTAGGCACATCTAAG-GCAAGAC----T 103
SotuI-1     CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-GAAATAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATTAGGCACATCTAAG-GCAAGAC----T 103
SotuI-4     CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-GAAATAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATTAGGCACATCTAAG-GCAAGAC----T 103
SotuI-15    CAAAT-GTAAAGT------AGGAATAGTAT-AGGACAGATGTAAA-----GTA-GAAATAGTATAGCATCTCTACTTGTATAAATAGA-GAGCCATTAGGCACATCTAAG-GCAAGAC----T 103
SotuI-2     CAGAT-GTAGAGT------AAGAATAGTAT-AGGACAGATGTAAA-----GGA-GAAATC------TACATTCTCCTCTATAAATAGG-AAGCCATTTAGGTAATCTAAG-GCAAGAC----T 97
SotuI-5     CAGAT-GTAAAGT------AAGAATAGTAT-AGGACAGATGTAAA-----GGA-GAAATC------TACATTCTCCTCTATAAATAGG-AAGCCATTTAGGTAATCTAAG-GCAAGAC----T 97
SotuI-10    CAGAT-GTAAAGT------AAGAATAGTAT-AGGACAGATGTAAA-----GGA-GAAGTC------TACATTCTCCTCTATAAATAGG-AAGCCATTTAGGTAATCTAAG-GCAAGAC----T 97
SotuI-7     CAGAT-GTAAAGT------AAGAATAGTMT-AGGACAGATGTAAA-----GGA-GAAATT------TACATTCTCCTCTATAAATAGG-AAGCCATTTAGGTAATCTAAG-GCAAGAC----T 97
SotuI-16    CAGAT-GTTAAGT------AAGGATAGTATTAGGACAGATGTTAA-----GGA-GGAATC------TACATTCTCCTCTATAAATAGG-AAGCCATTTAGGTAATCTAAG-GCAAGAC----T 98
Desiree-1   CAAGTAGTAAAGGCCA---ATGAATAGTAG-AAGGCAAGTGTAAACA---GTA-GGAGT-------CATTTATTTTCCTATATATAGG-CATAGATGTAAACATTGTAAA-GCRAG-C----C 101
Desiree-2   TGAATAGTAAAGGCCA---ATGAATAGTAG-AAGGCAAGTGTAAACA---GTA-GGAGT-------CATTTATTTTCCTATATATAGG-CATAGATGTAAGCATTGTAAA-GCAAG-C----C 101
Sotu6118    TGAATAGTAAAGGCCA---ATGAATAGTAG-AAGGCAAGTGTAAACA---GTA-GGAGT-------CATTTATTTTCCTATATATAGG-CATAGATGTAAACATTGTAAA-GCAAG-C----C 101
TPV         TGAACAGTAAAGGTCGTTCATGAATAGTAG-GAGTCATTTGTAAACA---GTA-AGAGT-------CG-TTTTAATTTTCTTTATA-------TAGAATGTAAATCTGAG-GAAGGAGGACAT 102
TVCV        TAAATAGTAAGAGTCGT-------TAGTAA-GAGTCGTTTGTGAATA---GTA-AGAGT-------CAATTTTTGTATTATAAATAG--CAGTTCAAATGTGAATAAAAA-ACAGGCTGCAGT 101
TraeI       CAAGGTACAAGATCGT---AGCTAAGGTAACAAGCTAAACGTCAAAGTCCACACGGAACTACTAGAGAGACTGACSTCTCTATGCAAAACATACAATAAGCAAACGTGAGTGCAAATGTACCT 237
                                    ****     * *   *** **      * *   * *           *       * *  ***                  *   *      *

SotuI-14    TTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCA--ATAAAATATC-AAGTTTCCAATCTAAGCTATGGATCAAAA------------------TATTCCAGATATCAC 199
SotuI-17    TTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCA--ATAAGATATC-AAGTTTCCAATCTAAGCTATGGATCAAAA------------------TATTCCAGATATCAC 199
SotuI-9     TTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCA--ATAAAATATC-AAGTTTCCAATCTAAGCTATGGATCAAAA------------------TATTCCAGATATCAC 199
SotuI-12    TTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCA--ATAAAATATC-AAGTTTCCAATCTAAGCTATGGATCAAAA------------------TATTCCAGATATCAC 199
SotuI-11    TTCTGGACGGAAWGCAAGCCTCC-TTGTACACCCCAATA--CTCA--ATAAAATATC-AAGTTTCCAATCTAAGCTATGGATCAAAA------------------TATTCCAGATATCAC 199
SotuI-21    TTCTSGACGGWAAGCAAGCCTCC-TTGTAAACCCCAATA--CTCA--ATAAAATATC-AAGTTTCCAATCTAAGCTATGGATCAAAA------------------TATTCCAGATATCAC 199
SotuI-8     TTCTCGACGGAAAGCAAGCCTCC-TTGTAAACMRRAATA--CTCA--ATAAAATATC-AAGTTTCCAATCTAAGCTATGGATCRAAA------------------TATTCCAGATATCAC 199
SotuI-1     TTCTCGACGGAAAGCAAGCCTCC-TTGTAAACAAAAATA--CTCA--ATAAAATATC-AAGTTTCCAATCTAAGCTATGGATCAAAA------------------TATTCCAGATATCAC 199
SotuI-4     TTCTCGACGGAAAGCAAGCCTCGGTTGTAAACACCTATA--CTCA--ATAAAATATG-GAGTTTCCAATCTAAGCTATGGATCAGAA------------------TATTCCAGATATCAC 200
SotuI-15    TTCTCGACGGAAAGCAAGCCTCC-TTGTAAACCCCAATA--CTCA--ATAAAATATC-AAGTTTCCAATCTAAGCTATGGATCAAAA------------------TATTCCAGATATCAC 199
SotuI-2     TTCCCGACGGAAAGCAAGCCTCT-TTGTAAACAAAAATATTCTCA--ATAAAATATCAAAGTTTTCAAGCTAAGTTATGAATCAAAATACAGTTATGGATTAAAATACACCAGATATCGA 214
SotuI-5     TTCCCGACGGAAAGCAAGCCTCT-TTGTAAACAAAAATATTCTCA--ATAAAATATCAAAGTTTTCAAGCTAAGTTATGAATCAAAATACAGTTATGGATTAAAATACACCAGATATCGA 214
SotuI-10    TTCCCGACGGAAAGCAAGCCTCT-TTGTAAACAAAAATATTCTCA--ATAAAATATCAAAGTTTTCAAGCTAAGTTATGAATCAAAATACAGTTATGGATTAAAATACACCAGATATCGA 214
SotuI-7     TTCCCGACGGAAAGCAAGCCTCT-TTGTAAACAAAAATATTCTCA--ATAAAATATCAAAGTTTTCAAGCTAAGTTATGAATCAAAATACAGTTATGGATTAAAATACACCAGATATCGM 214
SotuI-16    TTCCCGACGGAAAGCMAGCCTCT-TTGTAAACCAAAATATTCTCC--ATAAAATWTCCAAGTTTTCCAGGTTAGTTWTGGATCCAAAWACCGTTATGGAWTTAAATACCCCCGAWWTCGG 215
Desiree-1   TTCNCTACG--AAGCAAGCTCAT-TTGTAAA----AACACTCTTA--AATATATTTG-CATTTTAGAAACCTAATGGAGCCACTAGATAAATCAGATATCATGGAGAAGCCAGAGGCCA- 210
Desiree-2   TTCACTACG--AAGCAAGCTCRY-TTGTAAG----AACACTCTTA--AATATATTTG-CATTTTAGAAACCTAATGGAGCAACTAGATAAATCAGATATCATGGAGAAGCAAGAGGCAA- 210
Sotu6118    TTCACTACG--AAGCAAGCTCAT-TTGTAAA----AACACTCTTA--AATATATTTG-CATTTTAGAAACCTAATGGAGCAACTAGATAAATCAGATATCACGGAGAAGCAAGAGGCAA- 210
TPV         CCTCAGACATCC--TCATCCTCACCTTCTCTCTCTTATCTTCTCTCAATAAAATATCTGATTATATAC---AAACTTCTGAAAGCTATGGAGC---------ATTCAAACGAGTTACAAA 208
TVCV        TTTCAAGCATCCAACAATTCCTCTCTCTTCTCTCTAATATATTTT-----GCAGATATAAGCATACGA---AAGCTATGGAACAACAAGGAG----------ATAGCAGAAATCTACAAA 203
TraeI       AGCACGACTTACATCAGAACTAT-CTACATATGC-ATCAGTATCA-ACAAAGGGGTGGTGGAGTTTAACTGCARCAAGCTAGCTTTGACTCGGTG-------GCTAACCTGAACTACGAT 347
                   *        *        *          *     *          *  *   *  *        *             *

SotuI-14    A-CATGTATATCTATTTATGATTATGCATAGCTAA-TCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAAG-TATGTTTT 316
SotuI-17    A-CATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAAG-TATGTTTT 317
SotuI-9     A-CATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAAG-TATGTTTT 317
SotuI-12    A-CATGTATATCTATTTATGATTATGCATAG-TAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAAG-TATGTTTT 316
SotuI-11    A-CATGTATATCTATTTATGGTTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAAG-TATGTTTT 317
SotuI-21    A-CATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAAG-TATGTTTT 317
SotuI-8     A-CATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAAG-TATGTTTT 317
SotuI-1     A-CATGAATATCTATTAATGATTATGC-TATTTTTATCTGTTAATTCTTGAATATCATCAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAAG-TATGTTTT 316
SotuI-4     A-CATGTATATCTATTTATGATTATGCATAGCTCAATCTGTTAATTCTTGAATATCATTAGCATGCTTAAATAAGTTTATACTAGTTATTATATACTAGAATTATTTAAAG-CATGTTTT 318
SotuI-15    A-CATGTATATCTATTTATGATTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGCTTAAATAAGTTTATACTAGTTATTATATACTAGAATTATTTAAAG-CATGTTTT 317
SotuI-2     A-CATGTATATCTGTCTATGATTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAAG-CATGTTT- 331
SotuI-5     A-CATGTATATCTGTCTATGATTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAAG-CATGTTT- 331
SotuI-10    A-CATGTATATCTGTCTATGATTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAAG-CATGTTT- 331
SotuI-7     A-CATGTATATCTGTCTATGATTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAAG-CATGTTT- 331
SotuI-16    A-CMAGGAWATCYGTTTAWGATTAAGCATAGCTWAATTTATTGATTCCTTAATATCATTAGCATGATTGGATAAGTTTATATTAGTTATTATAWACTAGATTAATTCAAAG-CATGTTT- 332
Desiree-1   ---AGGTATATTGTTTATTAAATATGCAGAAGTAAATTCATATATTCCTGAATATCATAAATATGATTGAATTAGTTTATGCTAGGTATTATGGATTAGAATTATACGAAATCATGCTTT 327
Desiree-2   ---AGGTATATTGTTTATTAAATATGCAGAAGTAAATTCATATATTCCTGAATATCATAAATATGATTGAATTAGTTTATGCTAGGTATTATGGATTAGAATTATACGRAATCATGCTTT 327
Sotu6118    ---AGGTATATTGTTTATTAAATATGCAGAACTAAATTCATATATTCCTGAATATCATAAATATGATTGAATTAGTTTATGCTAGTTATTATGTATTAGAATTATACGAAATCATGCTTT 327
TPV         ACCAGGTAAGTTTATTTATAATCATGTTTAACTAAACTCTTATATTCCTTATAATCTCT-TGAATATCATAATTGTTTAT------CATGTTATAGTACTGTTATAAAGAA-CATCTTGA 320
TVCV        AACAGGTATATCATTTTCCAACTATGAACAGCTAAATTACTATATTCCTGATAATTTCTATGTAAATTATAATTGTTCAT------CATACTATAGTACTGTTATAAAAAT-CATCTTGA 316
TraeI       GCTATGTA-ACTCTTTTGTGGTGRTGCACA--CGAGTCCACATATTCGCCA----TATCAATACAMANATATGAATCCGCTSCCGTCTCCATACGAAAAARCCATCCATAG-CACTCACG 459
                 *    *   *        * *   *  *       *  ****                   *   *   *** **       **  *  * **   * **    *   **  *

SotuI-14    -CTAGTTTGC-TAACATGAGGAAAAAAAGAAAAACTT--CCAAAACTATGTCA----TCCTAAAGT-TGAAAC 380
SotuI-17    -CTAGTTTGC-TAACATGAGGAAAAAAAGAAAAACTT--CCAAAACTATGTCA----TCCTAAAGT-TGAAAC 381
SotuI-9     -CTAGTTTGC-TAACATGAGGAAAAAAAGAAAAGCTT--CCAAAACTATGTCA----TCCTAAAGT-TGAAAC 381
SotuI-12    -CTAGTTTGC-TAACATGAGGAAAAAAAGAAAAGCTT--CCAAAACTATGTCA----TCCTAAAGT-TGAAAC 380
SotuI-11    -CTAGTTTGC-TAACATGAGGACAAAAAGAAAGACTT--CCAAAACTATGTCA----TCCTAAAGT-TGAAAC 381
SotuI-21    -CTAGTTTGC-TWACATGAGGAAAAAAAGAAAGACTT--CCAAAACTATGTCA----TCCTAAAGT-TGAAAC 381
SotuI-8     -CTAGTTTGC-TAACATGAGGAAAAAAAGAAARACTT--CCAAAACTATGTCA----TCCTAAAGT-TGAAAC 381
SotuI-1     -GTAGTTTGC-TTACATGAGGAAGAAAAGATATACTKT-CCAAAAATAGTTCA----TCCTTAAGT-TGAAAC 381
SotuI-4     -CTAGTTTGT-TWATATGAGGAAAAAAGGARAGATTT--CCAAAACTATTCCA----TCCTAAAGT-TGAWAC 382
SotuI-15    -CTAGTTTGT-TAATATGAGGAAAAAAGGAAAAATTT--CCAAAACTATGCCA----TCCTAAAGT-TGACAC 381
SotuI-2     -ATAGTTTGC-TAACATGAGGAAAAAAGGAAAAACTT--CCAAAACTATGCCA----TCCTAAGGT-TGAAAC 395
SotuI-5     -ATAGTTTGC-TAACATGAGGAAAAAAGGAAAAACTT--CCAAAACTATGCCT----TCMWAAGGTGTGAAAC 396
SotuI-10    -ATAGTTTGC-TAACATGAGGAAAAAAGGAAAAACTT--CCAAAACTATGCCA----TCCTAAGGT-TGAAAC 395
SotuI-7     -ATAGTTTGC-TAACATGAGGWMAAAAGGAAAGACTT--CCAAAACTATGCCA----TCCTAAGGT-TGATAC 395
SotuI-16    -ATAGTTTGC-TAACATGAGGAAAAAAGGAAAAACTT--CCBAGACTATGCCA----TCCTAAGGT-TGAWAC 396
Desiree-1   -CTAGTTTAT-TAGCAAACAGGAAAAGCGAGAAACCTT-CCATAACTATGCCA----TCCTAAAGT-TGAAAC 392
Desiree-2   -CTAGTTTAT-TAGCAAACAGGAAAAGCGAGAAACCTT-CCATAACTATGCCA----TCCTAAAGT-TGAAAC 392
Sotu6118    -CTAGTTTAT-TAGCAAACAGGAAAAGCGAGAAACCTT-CCATAACTATGCCA----TCCTAAATT-TGAAAC 392
TPV         GAAAGTTTGC-CTGTCTAATAATGCTGAGAGTAGTTAAGCCCAGACTATGCCA----TCCTAAAGT-GGAAAC 387
TVCV        GAAAGTTTGC-CATGATAATA-TGCTAATAGTAGGTAAACCCAGACTATGCCA----TCCTAATGT-TGAAAC 382
TraeI       CTTATCTTGCGTACTTTAGAGTATCCACTWCCACATGTCTATGAACTATGCAAAGGGTCCGAGTTTCCATATM 532
               *****                     *         **   * **   *     **  **  *  ** **
```

Figure 3.6. Alignment of potato clones (Sotu) together with two additional PCR amplified sequences Desiree-1 and Desiree-2. TPV, TVCV and TraeI as for figure 3.3. The base substitutions to this new group are marked with grey.

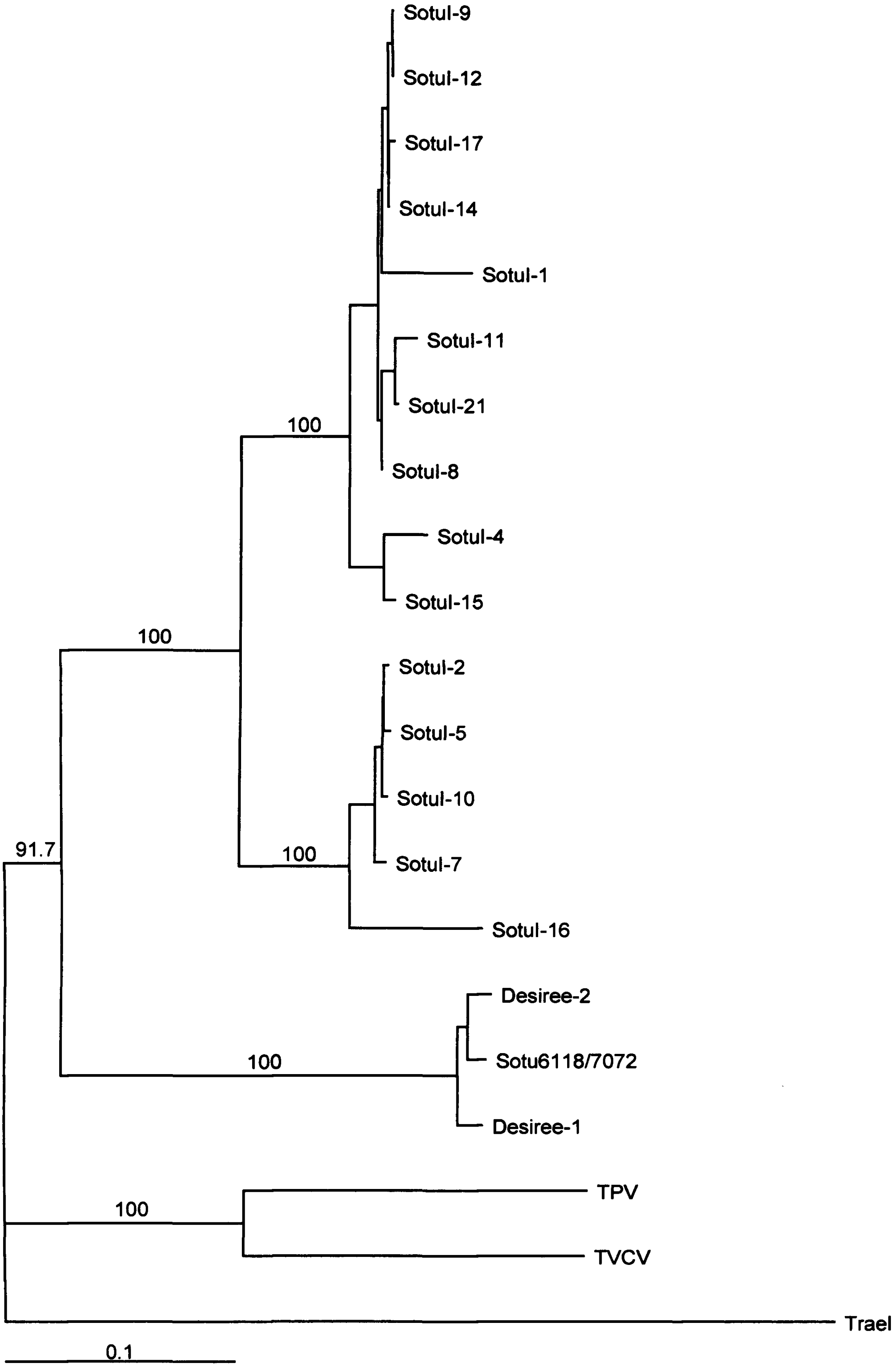


Figure 3.7. Phylogram from alignment of potato clones (Sotu) including two directly sequenced PCR products Desiree-1 and Desiree-2, as seen in figure 3.5. The origin of clone names are in table 3.1. Numbers on major branches are the bootstrap values in percentages based on 1000 trials. Bar indicates the branch length corresponding to a 10% difference.

Table 3.3. Similarities in percentages between individual clones and sequences from the alignment (figure 3.6). The different groups on the phylogram (figure 3.7) have been given different colours.

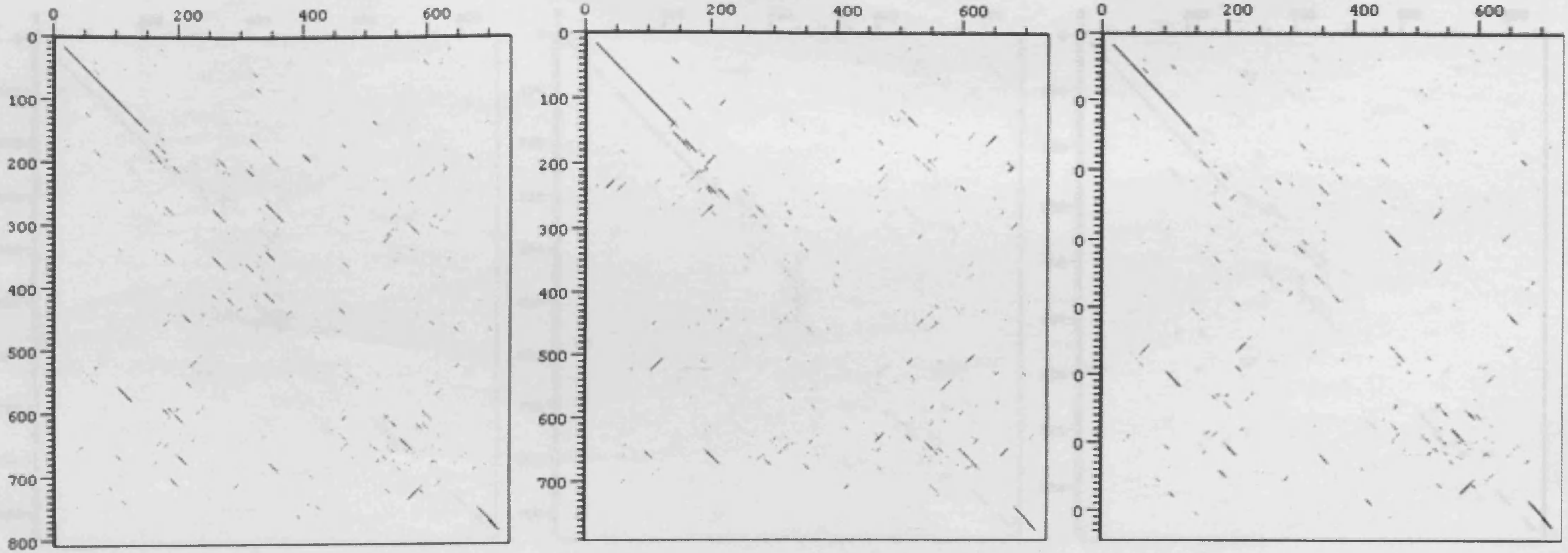
	Sotul-9	Sotul-12	Sotul-17	Sotul-14	Sotul-1	Sotul-11	Sotul-21	Sotul-8	Sotul-4	Sotul-15	Sotul-2	Sotul-5	Sotul-10	Sotul-7	Sotul-16	Desiree-1	Desiree-2	Sotu6118	TPV
Sotul-9																			
Sotul-12	100																		
Sotul-17	99	99																	
Sotul-14	99	98	99																
Sotul-1	95	94	95	94															
Sotul-11	97	97	97	97	93														
Sotul-21	97	97	97	97	93	97													
Sotul-8	98	98	98	98	94	97	97												
Sotul-4	93	93	93	93	91	93	93	93											
Sotul-15	96	96	96	95	91	95	95	95	96										
Sotul-2	87	87	87	87	83	85	85	86	84	86									
Sotul-5	86	86	86	86	82	84	85	85	83	86	98								
Sotul-10	87	87	87	87	83	85	85	86	84	86	99	98							
Sotul-7	86	86	86	86	82	85	85	86	84	86	97	97	97						
Sotul-16	79	79	79	79	75	78	78	79	77	79	89	88	89	88					
Desiree-1	67	67	66	67	64	65	65	66	66	66	66	67	68	67	63				
Desiree-2	68	68	67	68	65	66	66	67	67	67	67	67	69	67	63	96			
Sotu6118	68	68	68	68	65	66	66	67	68	68	67	68	69	68	64	96	97		
TPV	42	41	41	41	39	41	41	41	41	42	41	40	41	40	28	62	63	64	
TVCV	36	36	36	36	33	35	36	35	35	36	39	38	39	37	29	58	60	58	72

similarity with several other pararetroviruses. As examples, SotuIII showed a similarity of 85% to full length TPV and TVCV. The clone Sotu6118/7072 had a similarity of 66% to TPV and 70% to TVCV. SotuI-1 had a similarity of 60% to both TPV and TVCV. Apart from TPV and TVCV, only additionally TPV-like sequences and newly sequenced PRV-L fragments from *Nicotiana otophora* and *N. tomentosiformis* showed high similarity to the Sotu sequences. Figure 3.8 A and B shows dotplots of SotuI-1 and Sotu6118/7072 respectively against the most significant similar sequences from the database. The areas of very high similarity, a diagonal line from top left and bottom right covers about 150 bp from the beginning of SotuI-1 and 50 bp in the end of the sequence (figure 3.8 A). For Sotu6118/7072 the high similarity are maintained for longer, at least 250 bp from the beginning of the sequence and two sections of 50 bp at the end with several areas of similarity in the middle (figure 3.8 B). The alignment figure 3.4 also indicates that the sequences tend to have higher homology at the beginning and at the end.

3.3.3 Organization of PRV-L sequences

Dotplot is a tool to display homology between two sequences and highlight internal repeats (Pustell and Kafatos, 1982). One sequence is placed along the X-axis the other along the Y-axis, where there is a match between the two sequences a dot will be generated in the internal square. Two identical sequences will generate a diagonal line of 100% homology. To avoid getting too many dots (noise) a filtering system is used. Instead of comparing the sequences base by base a window of consecutive bases and a percentage of homology is chosen. For instance a dot is placed if a homology of 70% is found within 30 bases. Lines generated parallel to the diagonal indicates tandem repeats, lines perpendicular to the diagonal indicates inversions and lines crossing the diagonal indicate palindromes.

Dotplot analysis was carried out on the 8 kb TPV sequence with the program Megalign (figure 3.9). The most obvious result on the dotplot is the area with large tandem repeats seen as parallel lines to the diagonal between 6000 and 7000 bp. A noteworthy fact is the near absence of other dots/repeat sequences in this region; there are very few short lines in the area vertical and horizontal from the repeat region. This is the area of TPV, amplified with primer pairs TPV6241-TPV7041 and TPV6118-TPV7072 and which is found in very diverse plant families from both Gymnosperms and Angiosperms (chapters IV and V). A similar pattern is seen from

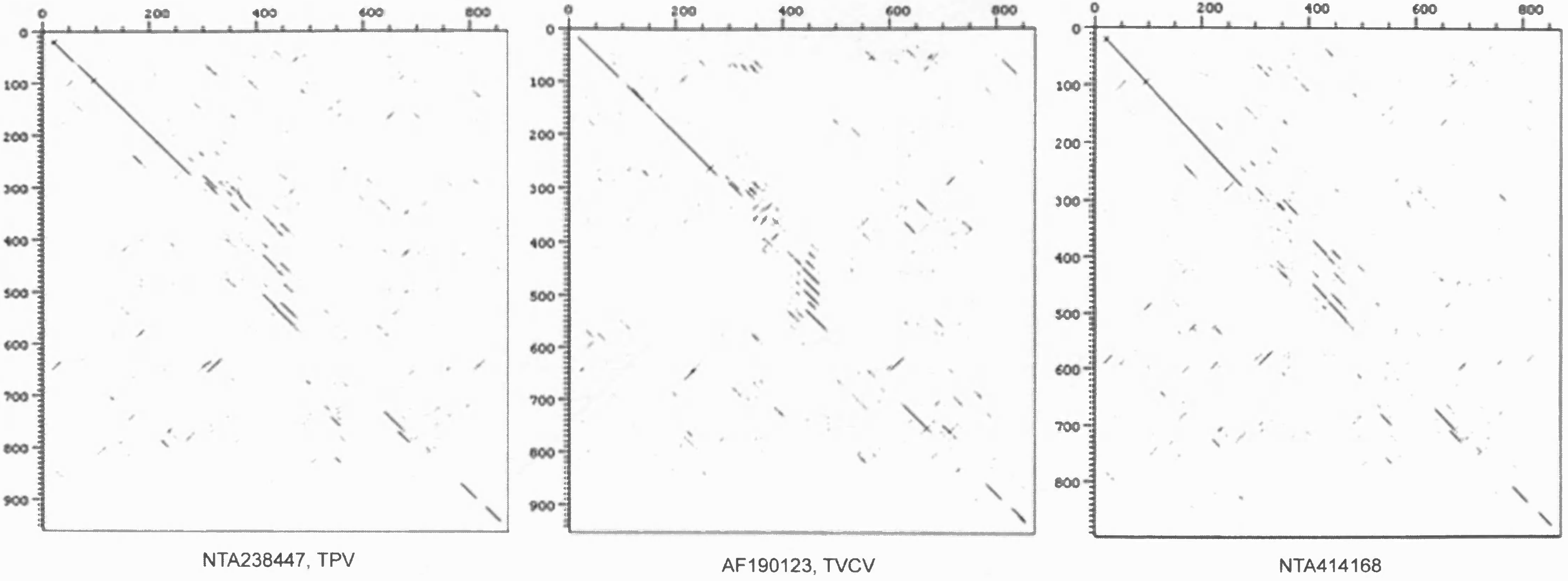


NTA238447, TPV

AF190123, TVCV

NTA414168

Figure 3.8 A. Dotplot of SotuI-1 (horizontal) compered to NTA238447 (TPV), AF190123 (TVCV) and NTA414168, a tabaccopararetrovirus-like sequence (Jacowitsch *et al.*, 1999) (vertical). The diagonal lines form top left and bottom right show sites of very high similarity. Dotter tool from Center for Genomics Research, Karolinska Institutet, greyramp tool settings 25/51.



NTA238447, TPV

AF190123, TVCV

NTA414168

Figure 3.8 B. Dotplot of Sotu6118/7072 (horizontal) compared to NTA238447 (TPV), AF190123 (TVCV) and NTA414168, a tobacco pararetrovirus-like sequenc (Jacowitsch *et al*., 1999) (vertical). The diagonal lines form top left and bottom right show sites of very high similarity. Dotter tool from Center for Genomics Reserach, Karolinska Institutet, greyramp tool settings 25/51.

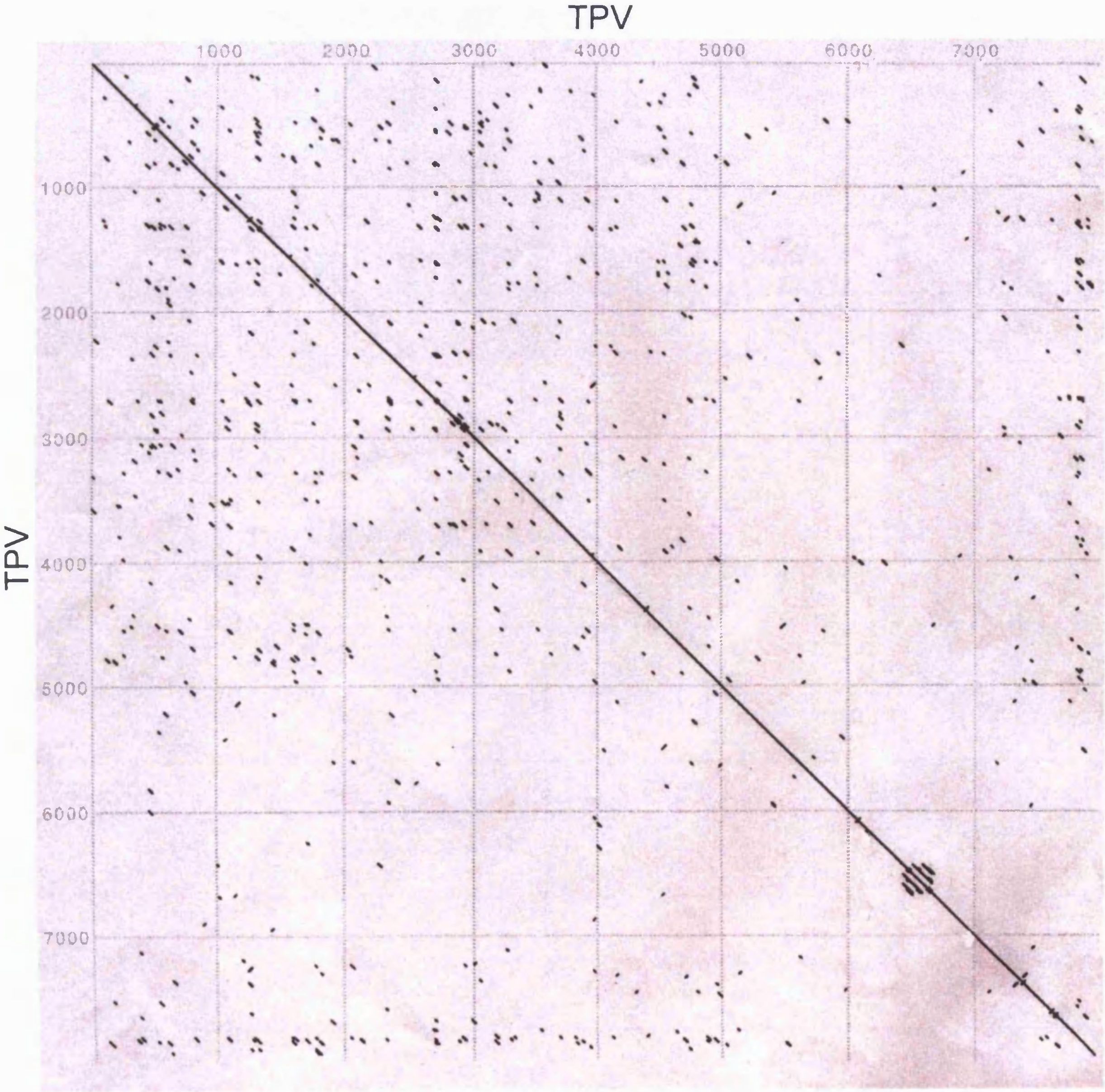


Figure 3.9. Dotplot of TPV (tobacco pararetrovirus). With 70% identity over 30 bases.

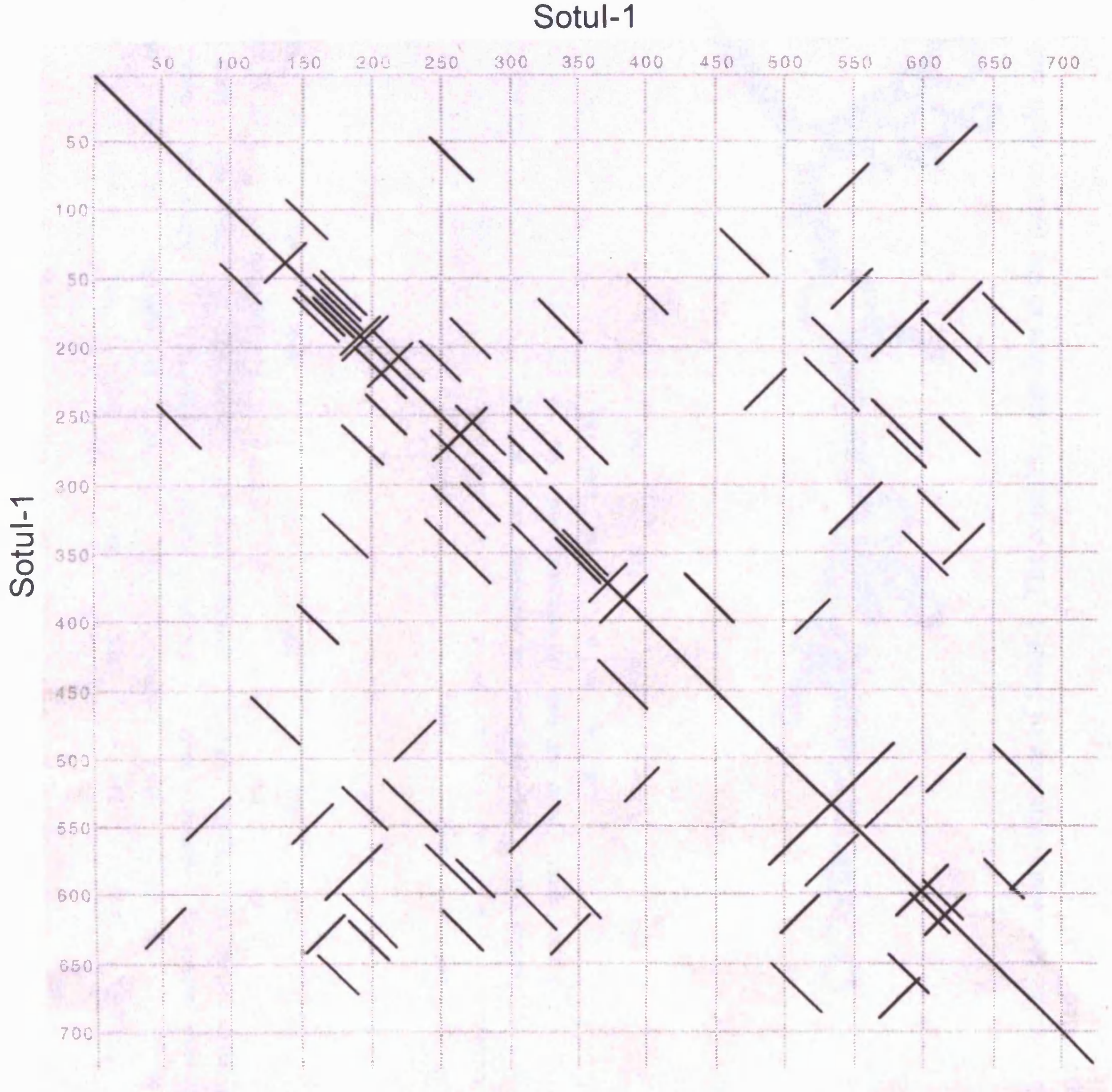


Figure 3.11. Dotplot of a potato clone Sotul-1. With 60% identity over 30 bases.

plt22jpg by D. Stewart and M. Zuker

©2002 Washington University

```
     310       320       330       340         350       360         370       380       390          400
.-UA|     UA    G      A  GUA      GAA    .-AUAG        U       .-AA     A   CAUCUCU     U   .-AA    GA-   A
    AAGUAA  AUUA GAGACA AU   AAGUAG   UAGU      GACAGAUG AAAGUAG    AUAGU UAG       ACUUG AUA    UAGA   GCC U
    UUCGUU  UGAU UUUUGU UG   UUUAUU   AUCA      UUGUCUAU UUUUAUC    UAUCG AUC       UGAAC UAU    AUCU   CGG U
\ --^     --    G      A  AAG      AAG    \ ----        -       \ --     A   UAACCUU     -   \ --    ACA   A
  650         640       630       620           560        550            490       480              410
```

```
120       130       140         150       160        170       180
  AGACAAUGUAAUUACAGC      G    .-A    A   C   C-  C        A   AAG  AAA
                    UGUCCA CUAG   AUAA AAG CAA  GA AAGAUAAA AGC   AU   A
                    ACAGGU GAUU   UAUU UUC GUU  UU UUCUGUUU UUG   UG   U
  A-----------------      -    \ -    A   A   UA  A        G   AAA  AAA
                       260          220       210       200       190
```

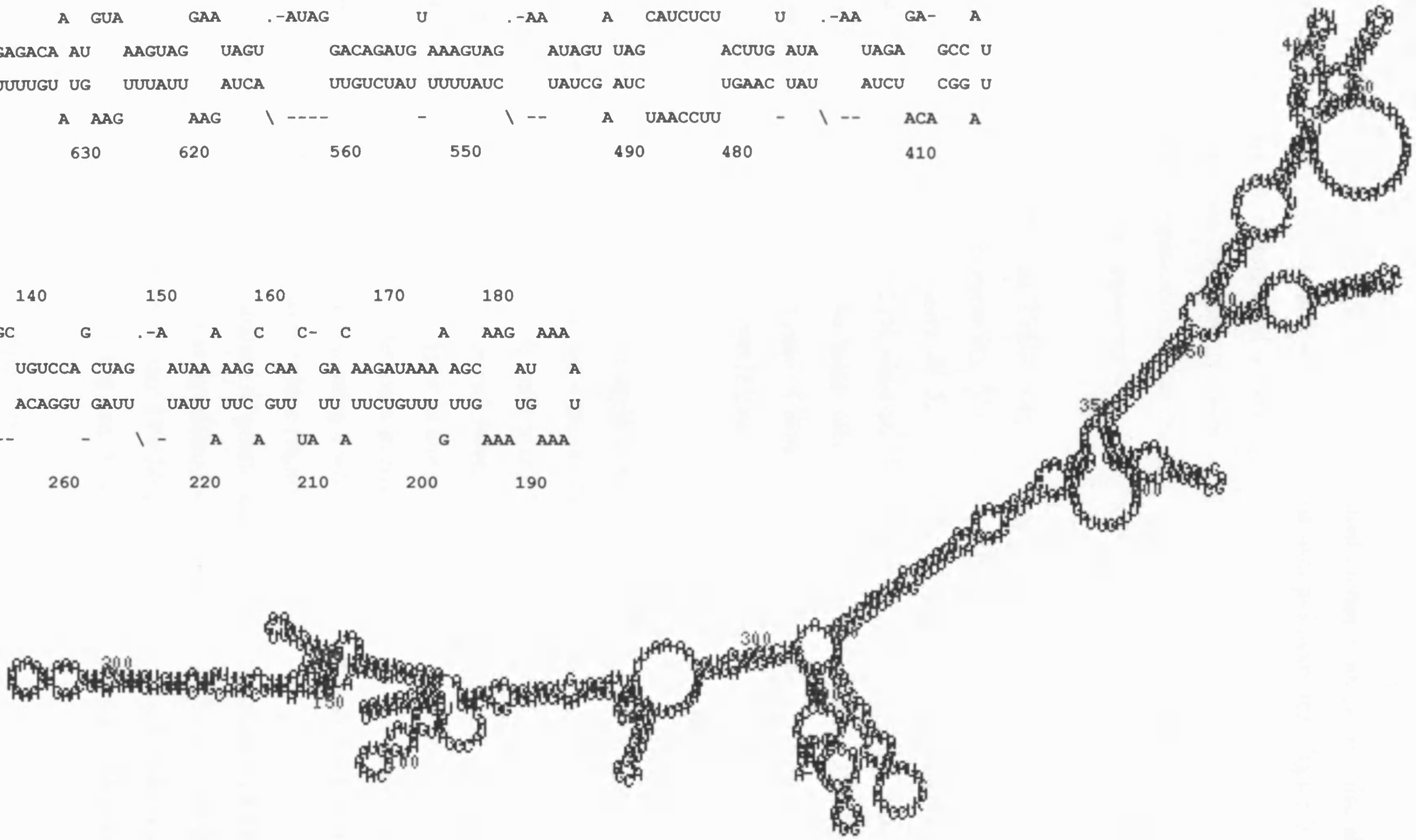


Figure 3.12. Possible secondary structure of SotuI-1. The complete sequence to the bottom right and two details at the top left. The programme mfold has converted T to U.

dotplot of TVCV (figure 3.10). Dotplots of individual clones from potato also show an extensive pattern of tandem repeats, inversions and palindromes (figure 3.11). Figure 3.12 shows an example of a secondary structure of SotuI-1 and how the inverted repeats can form stretches of double stranded DNA. The programme mfold is working with RNA sequences and has changed the T bases in the submitted DNA sequence into U. Similar secondary structures may arise from transcribed RNA sequences.

The region of TPV and TVCV homologous to the SotuI sequences had a GC content of 33% and 32% respectively. The clone SotuI-1 had a GC content of 31% and Sotu6118/7072 a GC content of 32%. TVCV in the region homologous to SotuIII-1 had a GC content of 25% whereas TPV had a GC content of 34%. SotuIII-1 had a GC content of 26%. So these sequences have a rather low GC content, strangely especially for the RT region of SotuIII-1, but consistent with a wide range of GC content found in retroelements (Figure 1.4, 28-62%).

3.3.4 Southern hybridization

Potato genomic DNA digested with eight different enzymes (figure 3.13A) was blotted to a membrane and hybridized with an AlkPhos labelled potato clone SotuI-12 (figure 3.13B). The hybridization gave many bands from high molecular weight to 1 kb over a general background smear. Strong bands was seen to *Eco*RI digest at about 3.2 kb and 5 kb and to *Xba*I digest at about 4.3 kb with a double band at about 2.2 kb. There were only minor differences in hybridization pattern between using a distinct clone as a probe compared to using a whole PCR product amplified with the same primers (TPV6241-7041) as the clone (figure 3.13C).

Another gel showing separation of potato genomic DNA digested with seven different enzymes (figure 3.14A) was hybridized with the clone SotuIII derived from PCR amplification with the primer pair TPV3462-4579 (figure 3.14B). This clone gave strong bands with *Eco*RI (2.3 kb) and *Xba*I (2.8 kb and 4.3 kb) and general smear with others (figure 3.14B).

The enzymes *Msp*I and *Hpa*II showed that the PRV-L sequences are methylated. *Msp*I and *Hpa*II (both 5' C↓CGG 3') depend on the degree of methylation. *Msp*I is inhibited by 5-methylcytosine at the 5' C residues but is not sensitive to the presence of 5- or 4-methylcytosine at the internal C residues. *Hpa*II is inhibited by any form of methylation at either C residue, implying that *Hpa*II is more

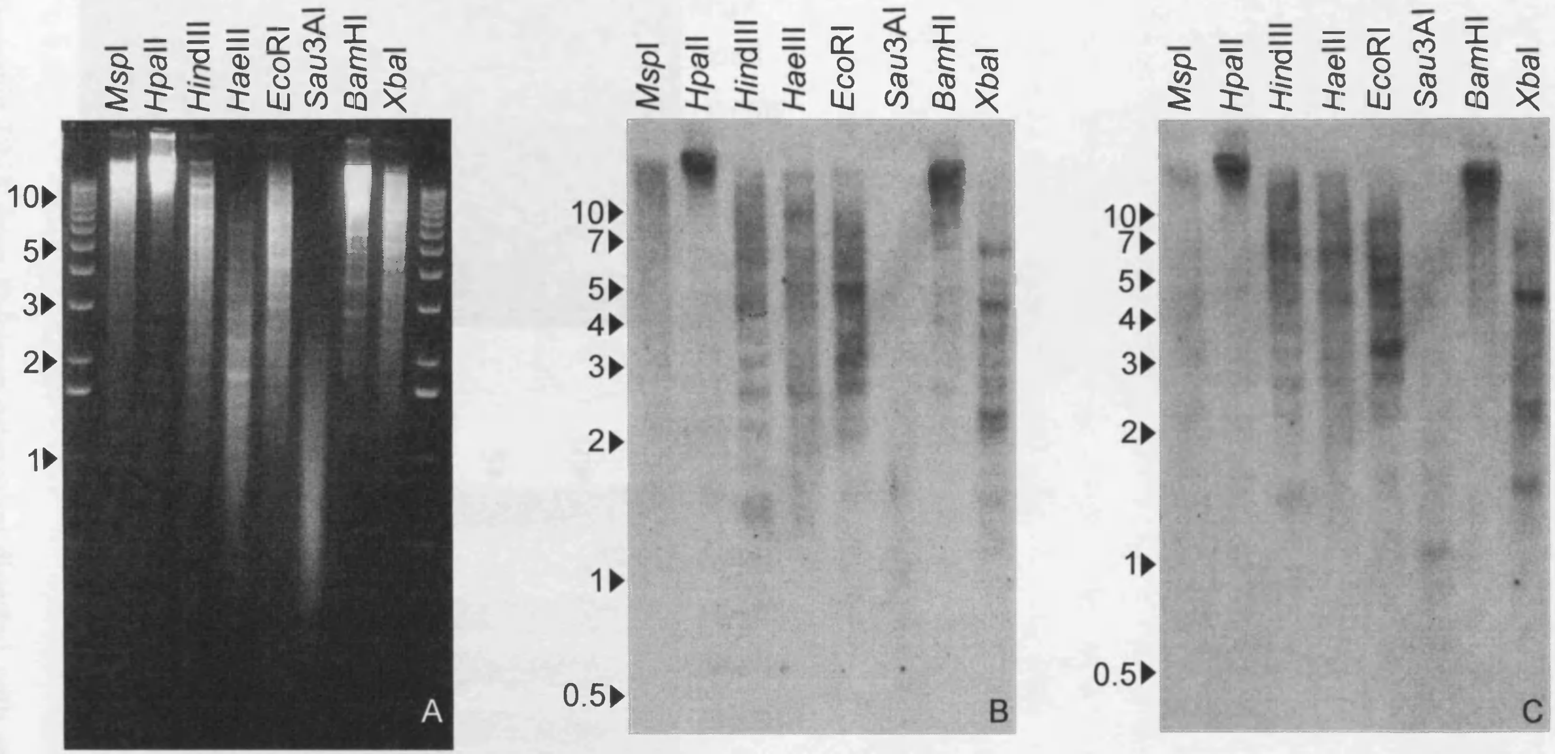


Figure 3.13. Genomic DNA from *Solanum tuberosum* digested with eight different enzymes, A) Ethidium bromide stained gel image. B) Southern blot hybridized with a *Solanum tuberosum* clone SotuI-12, C) and a whole PCR product . Numbers to the left of images are sizemarkers in kilobases.

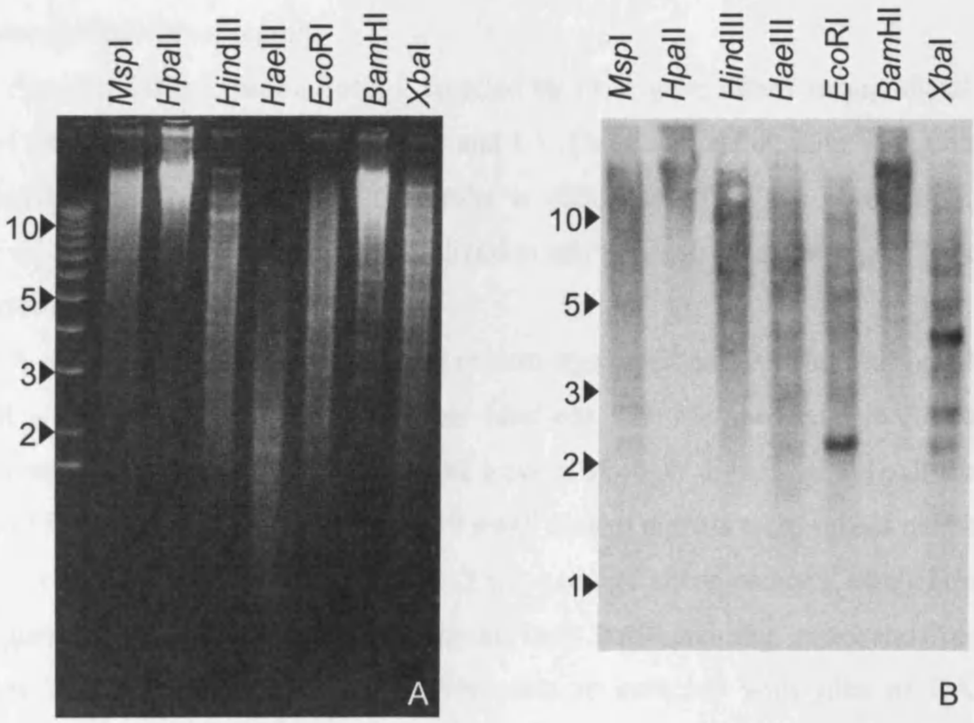


Figure 3.14. Genomic DNA from *Solanum tuberosum* digested with seven different enzymes. A) Ethidium bromide stained gel image. B) Southern blot hybridized with a *Solanum tuberosum* clone SotuIII-1. Numbers to the left of images are sizemarkers in kilobases.

sensitive to methylation than *Msp*I. *Hpa*II did not readily digest the potato DNA and hybridization of any probe was to high molecular weight DNA (figure 3.13 and 3.14).

3.3.5 *In situ* hybridization

As potato, *Solanum tuberosum*, has a low C-value and many chromosomes, the chromosomes appeared rather small. Because of the size and little chromosome morphology, identification of individual chromosomes or features such as the centromere was difficult. Starch grains clustering around cells were also a problem in some instances as they either covered the chromosomes or gave background fluorescence.

SotuIII-1 labelled by nick-translation hybridized to the potato chromosomes. In figure 3.15B and C appearing as 1-5 stronger single or double dots and several minor sites on at least 36 chromosomes, the remaining showing little or no signal. Figure 3.15D shows an enlargement of four homologous chromosomes with 5S signal (red), the hybridization of SotuIII-1 (green) was rather similar between the chromosomes. In figure 3.16 C and D SotuIII-1 showed less prominent signal but still hybridizing to nearly all chromosomes. The probe did not seem confined to a particular chromosomal region.

Another PRV-L probe SotuI-2 labelled by PCR gave rather strong signal to most of the chromosomes (figure 3.17B and C). There is considerable background hybridization but the majority of the probe is concentrated on the chromosomes. Nearly all chromosomes had some hybridization and in some cases the signal looked like centromeric bands or double dots.

A related probe SotuI-1 gave less intense signals (figure 3.18B (red) labelled by PCR and 3.19B (green) random primer labelled). The chromosomes in 3.18 are more contracted than they are in 3.19 and gave a stronger signal on at least three quarter of the chromosomes. In figure 3.19 small dots of signals were spread over the length of over half of the chromosomes. Two pairs of chromosomes were chosen from figure 3.19D (arrows) to show the signals from DAPI staining, *gypsy* and SotuI-1 (figure 3.20), signals from both probes seem to coincide with sites of DAPI brightness corresponding to areas of heterochromatin, the *gypsy* probe spreading to other areas also.

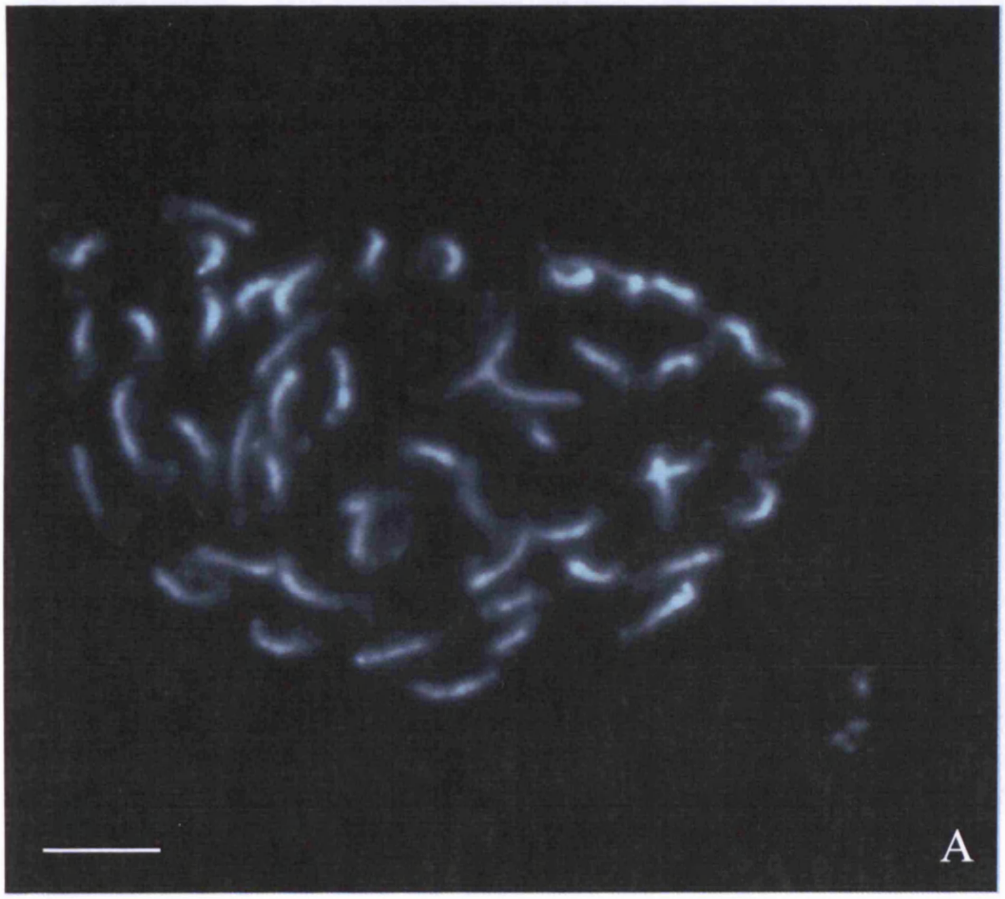


Figure 3.15. Metaphase spread of *Solanum tuberosum* 'Desiree'. A) 48 DAPI stained chromosomes. Bar represents 10 µm. *Continued next page.*

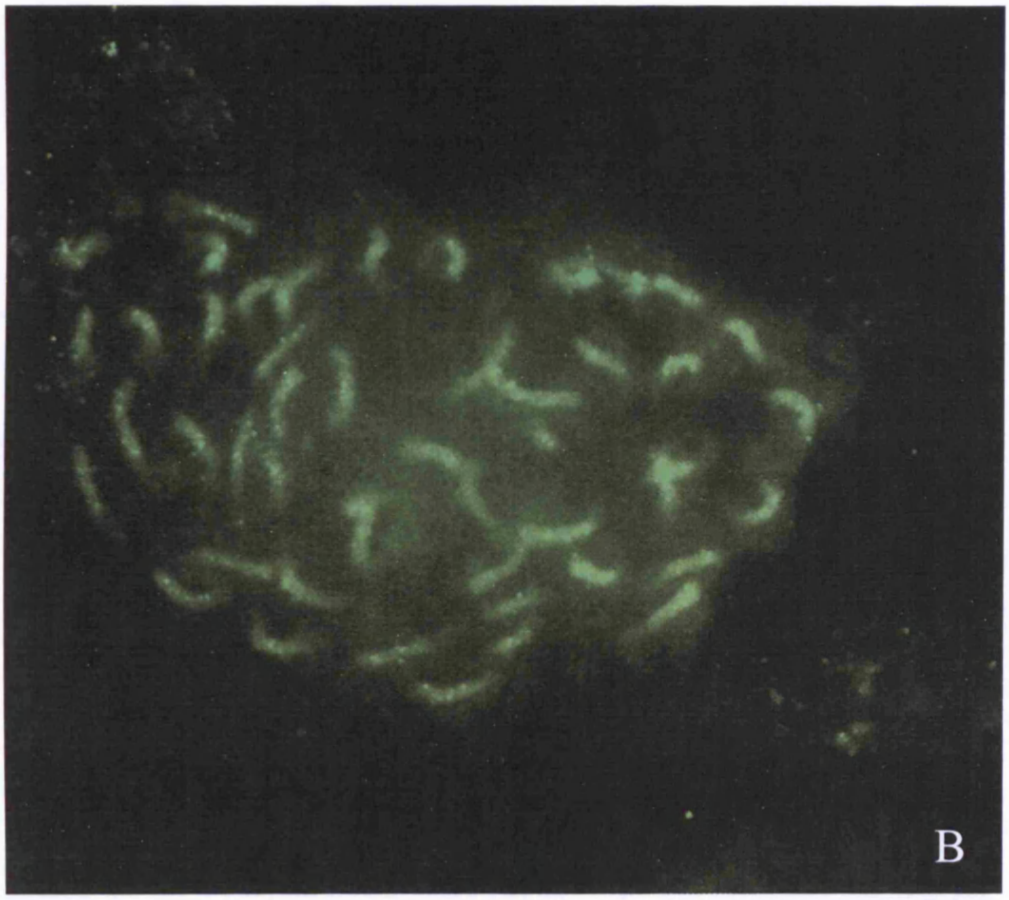


Figure 3.15 *continued.* B) *In situ* hybridization with a PRV-L clone, SotuIII-1 labelled and detected with digoxigenin/FITC. In B and C the PRV-L probe is shown to hybridize to all chromosomes, more strongly in some, dispersed with single and double dots along the chromosome arms. *Continued next page.*

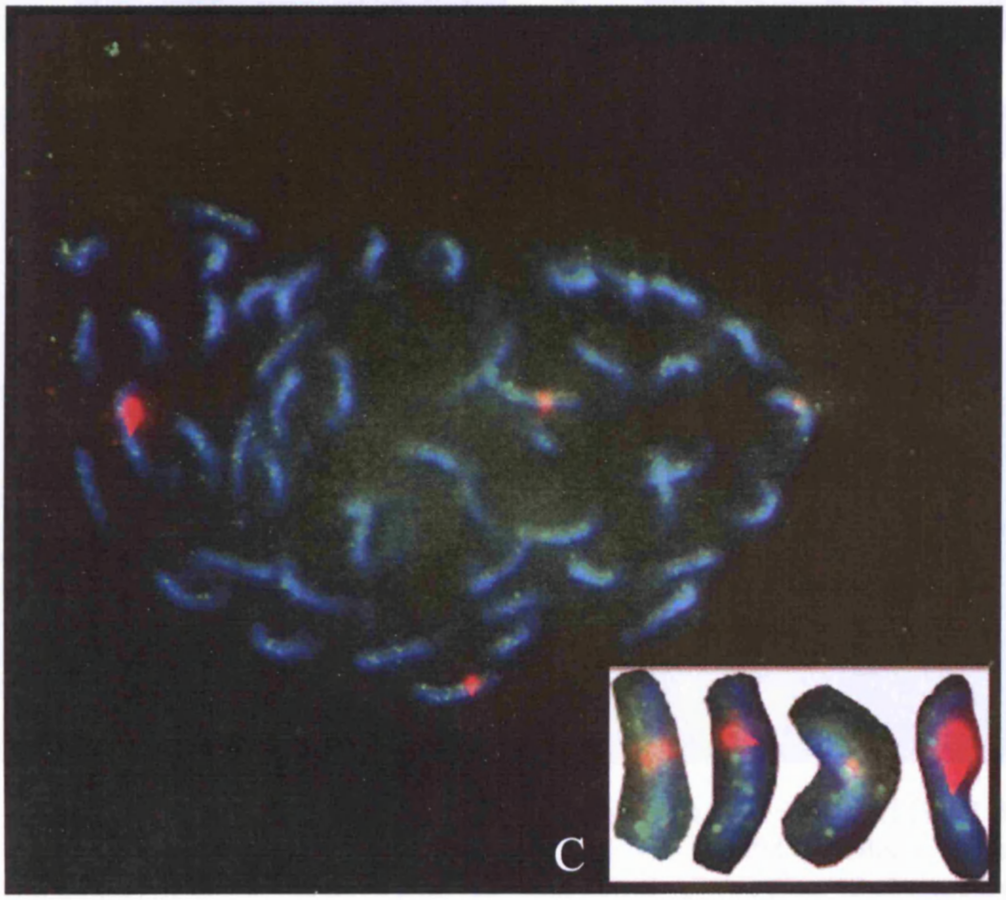


Figure 3.15 *continued.* C) Overlay of A and B including hybridization with 5S rDNA (pTa794) as a red site on four chromosomes. Insert, enlargement of the four 5S chromosomes, for two of them the red colour is overexposed forming a triangle.

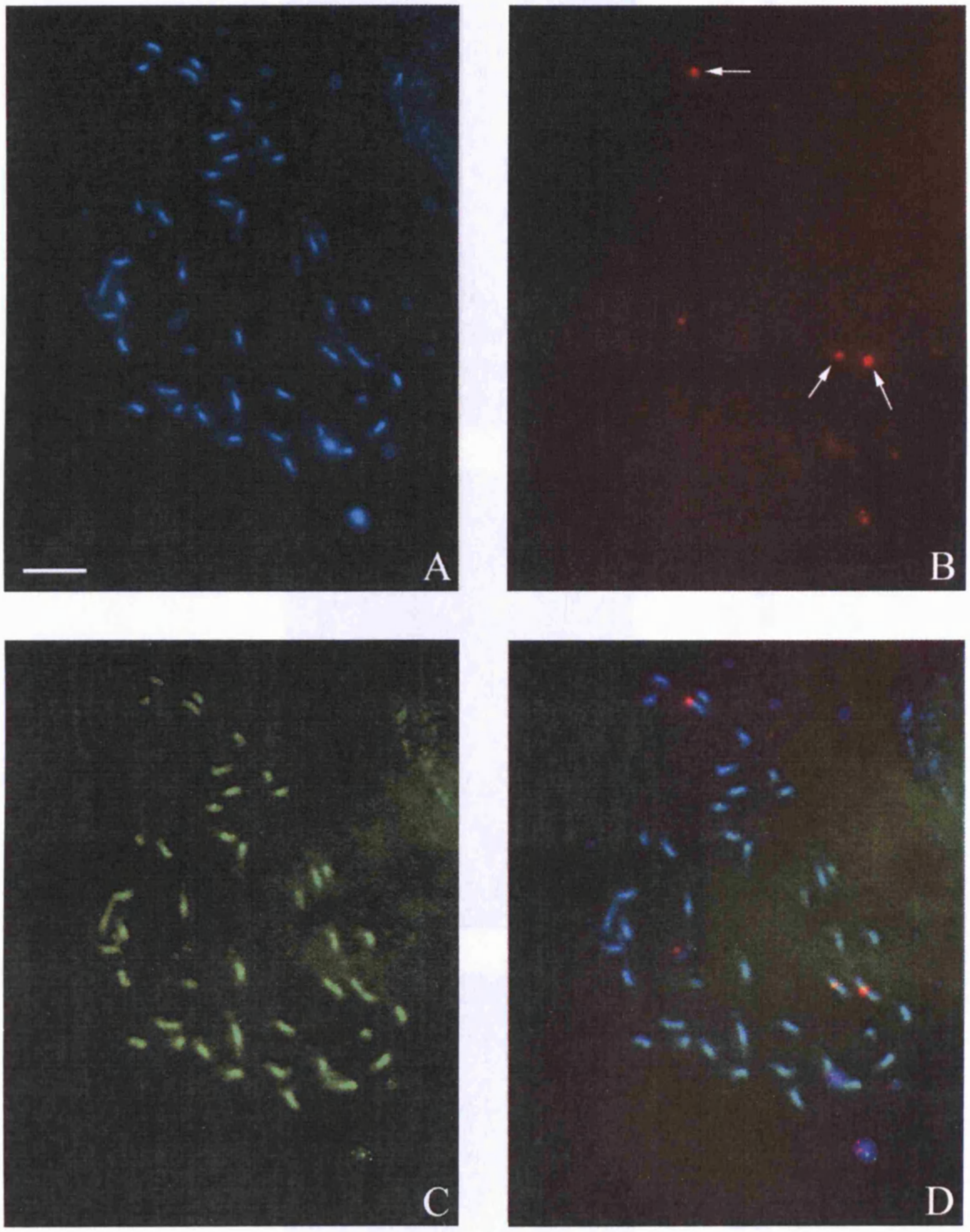


Figure 3.16. Incomplete metaphase of *Solanum tuberosum* 'Desiree'. A) DAPI stained chromosomes. B) *In situ* hybridization with 5S rDNA (pTa794) labelled and detected with biotin/Alexa594, the three sites are marked with arrows (one 5S chromosome is missing). C) *In situ* hybridization with a PRV-L clone SotuIII-1 labelled and detected with digoxigenin/FITC. D) Overlay of A, B and C showing three 5S sites and bands or dots on all chromosomes from SotuIII-1. Bar represents 10 μm.

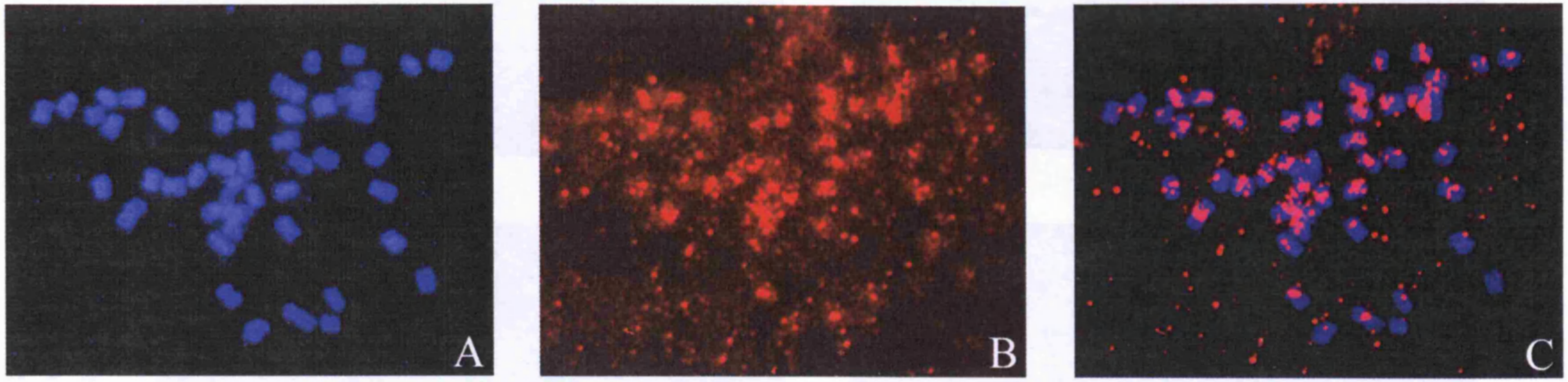


Figure 3.17. Metaphase of *Solanum tuberosum* 'Desiree'. A) DAPI stained chromosomes. B) *In situ* hybridization with a PRV-L clone, SotuI-2, labelled and detected with biotin/Cy3. C) Overlay of A and B. The probe hybridized strongly to the chromosomes, in some cases as centromeric bands or double dots.

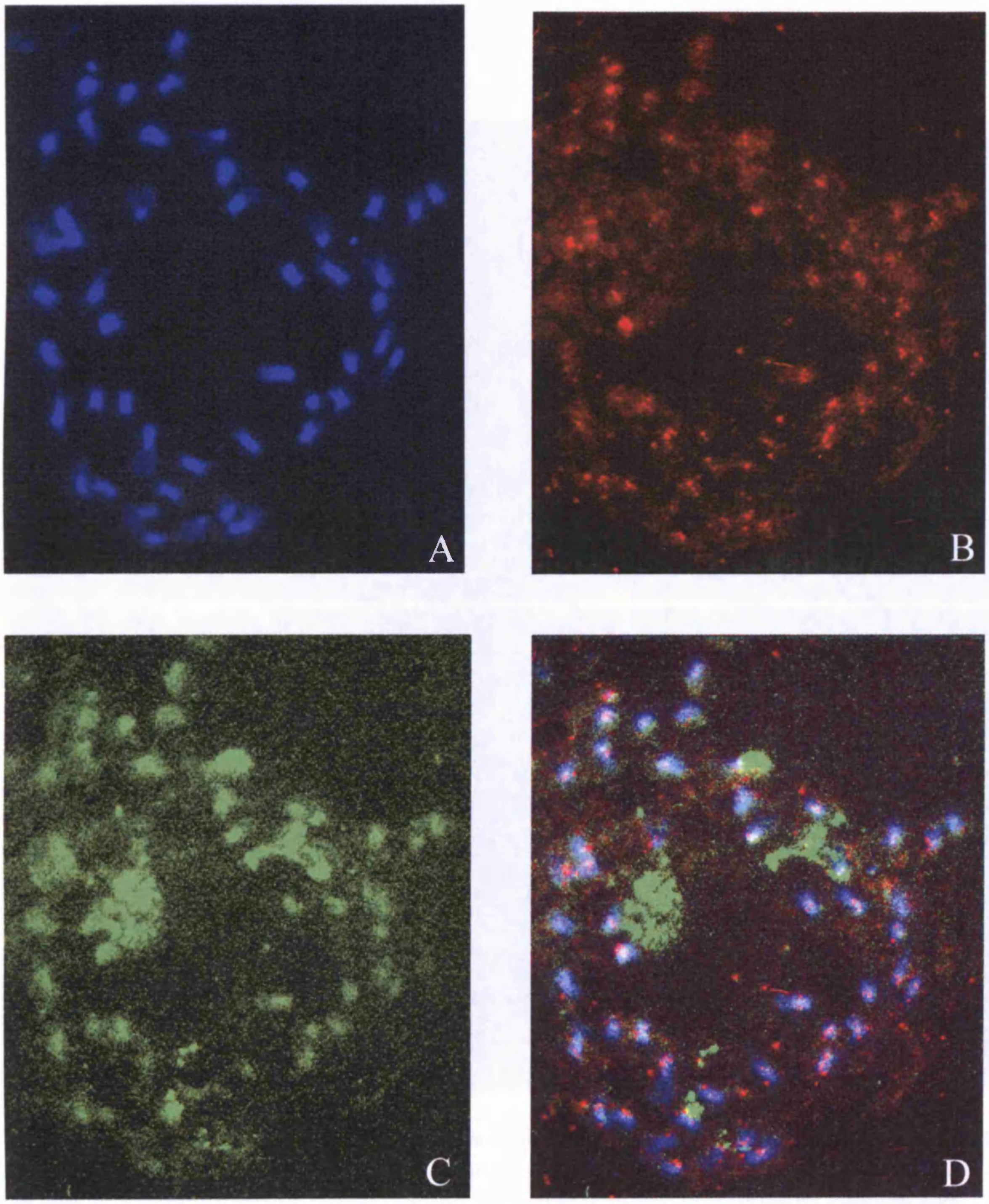


Figure 3.18. Metaphase of *Solanum tuberosum* 'Desiree'. A) 48 chromosomes stained with DAPI. B) *In situ* hybridization with a PRV-L clone (sotul-1) labelled and detected with biotin/ Cy3. C) *In situ* hybridization with a 45S rDNA probe (pTa71). D) Overlay of A, B and C showing the sites of two large and two small rDNA sites and the scattering of the virus probe with double dots on nearly all chromosomes.

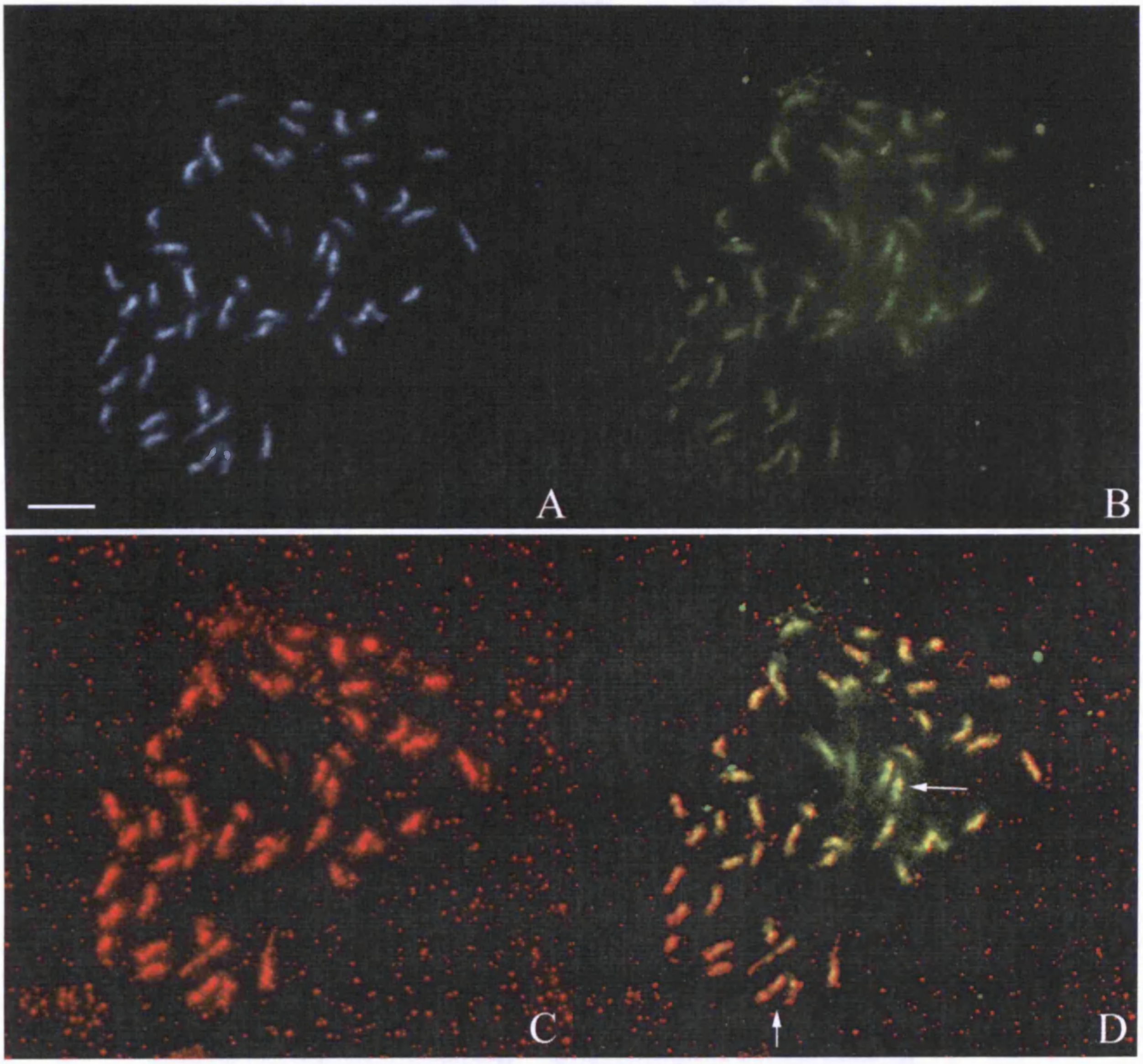


Figure 3.19. Metaphase of *Solanum tuberosum* 'Desiree'. A) 48 DAPI stained chromosomes. B) *In situ* hybridization with a PRV-L probe, Sotul-1, labelled and detected with digoxigenin/FITC. C) *In situ* hybridization with a potato specific *gypsy* retroelement PCR product labelled and detected with biotin/Alexa594. D) Overlay of A, B and C. The *gypsy* probe hybridized strongly to all chromosomes and along the entire length. It is an uncleaned PCR product and therefore also gives some background. Sotul-1 hybridized weakly along the length of most chromosomes. Figure 3.19 shows a comparison of DAPI stain, *gypsy* and Sotul-1 signal on two sets of chromosomes marked with arrows in D. Bar represents 10 μm.

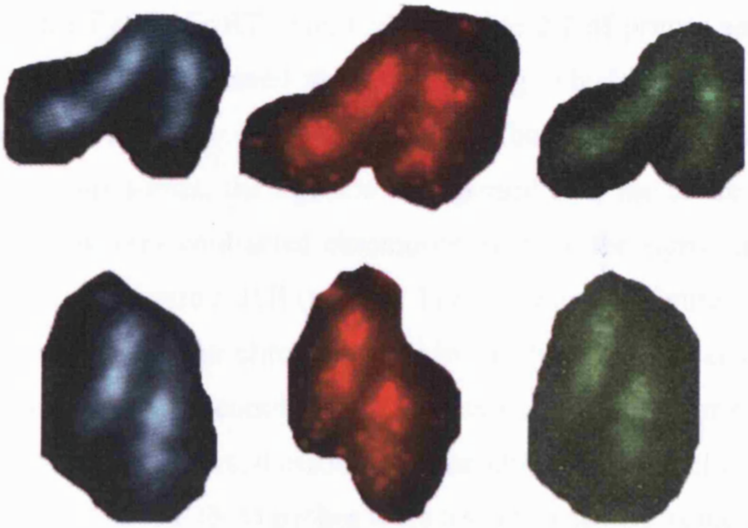


Figure 3.20. Pairs of chromosomes from figure 3.18 to compare the signals fromDAPI, *gypsy* and Sotul-1. The probes hybridize to heterochromatic areas corresponding to area of DAPI brightness.

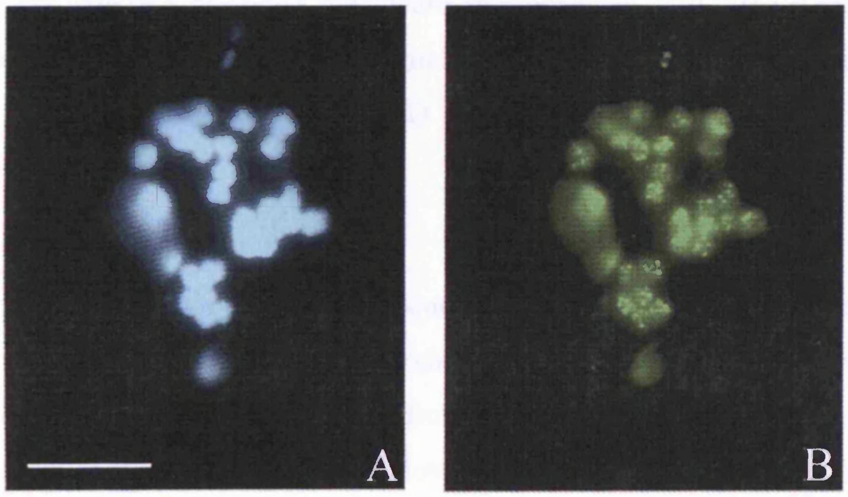


Figure 3.21. Part of a *Solanum tuberosum* 'Desiree' metaphase. A) Very contracted DAPI stained chromosomes. B) *In situ* hybridization with a potato specific *gypsy* retroelement PCR product labelled and detected with digoxigenin/FITC. Many distinct dots of signal are seen on the chromosomes. Bar represents 10 μm.

Potato specific *copia* and *gypsy* retroelement probes were synthesized by PCR amplification of genomic DNA with degenerate primers for *copia* and *gypsy* RT region (Tyl-1 and Tyl-2; GyRT1 and GyRT4, table 2.2 of primer sequences). The PCR products were not cleaned before labelling which probably resulted in background noise. The *gypsy* probe gave a lot of background but also hybridized strongly to all chromosomes, the signal was dispersed over the entire length (figure 3.19C (red)). A few very contracted chromosomes show the signal from the *gypsy* probe as distinct dots (figure 3.21B (green)). The *copia* probe (figure 3.22) gave dots of signals dispersed over the chromosomes. In the figure both cells with longish chromosomes (top) and very contracted chromosomes are seen. Some chromosomes had stronger signals than others, these varied in number from 5-12 chromosomes.

Ribosomal DNA (rDNA) probes were used to identify some chromosomes and are also a convenient positive test of the *in situ* hybridizations. The ones used were pTa794 (Gerlach and Dyer, 1980) and pTa71 (Gerlach and Bredbrook, 1979). The pTa71 probe hybridizes to the 45S rDNA of the nucleolar organizer region (NOR) and the pTa794 probe hybridizes to 5S rDNA at other regions of the genome. The 5S distinctively gave four signals, the probe hybridizing close to the centromere of the short arm of a large chromosome (figure 3.23), perhaps with some minor sites on other chromosomes. For 5S see also figure 3.15C, D and 3.16B, D. The 45S signals were often indistinct covering a larger area, as two major and one or two medium sites. Figure 3.24 shows metaphase chromosomes with three signals (B and C (red)) and an interphase with four signals (D and E (green)). Hybridization with pTa71 is also shown in figure 3.18C and D.

3.4 Discussion

In this chapter it is shown that PRV-L sequences are detected by PCR, Southern and *in situ* hybridization in the potato genome. Specific PRV-L primers are able to amplify fragments of the expected size from potato genomic DNA (figure 3.1, 3.2 and 3.3). Dotplot analysis showed that most of the primers are derived from an area of many repeat sequences (figure 3.9 and 3.10) and data also showed the sequences to be methylated and have a low GC content.

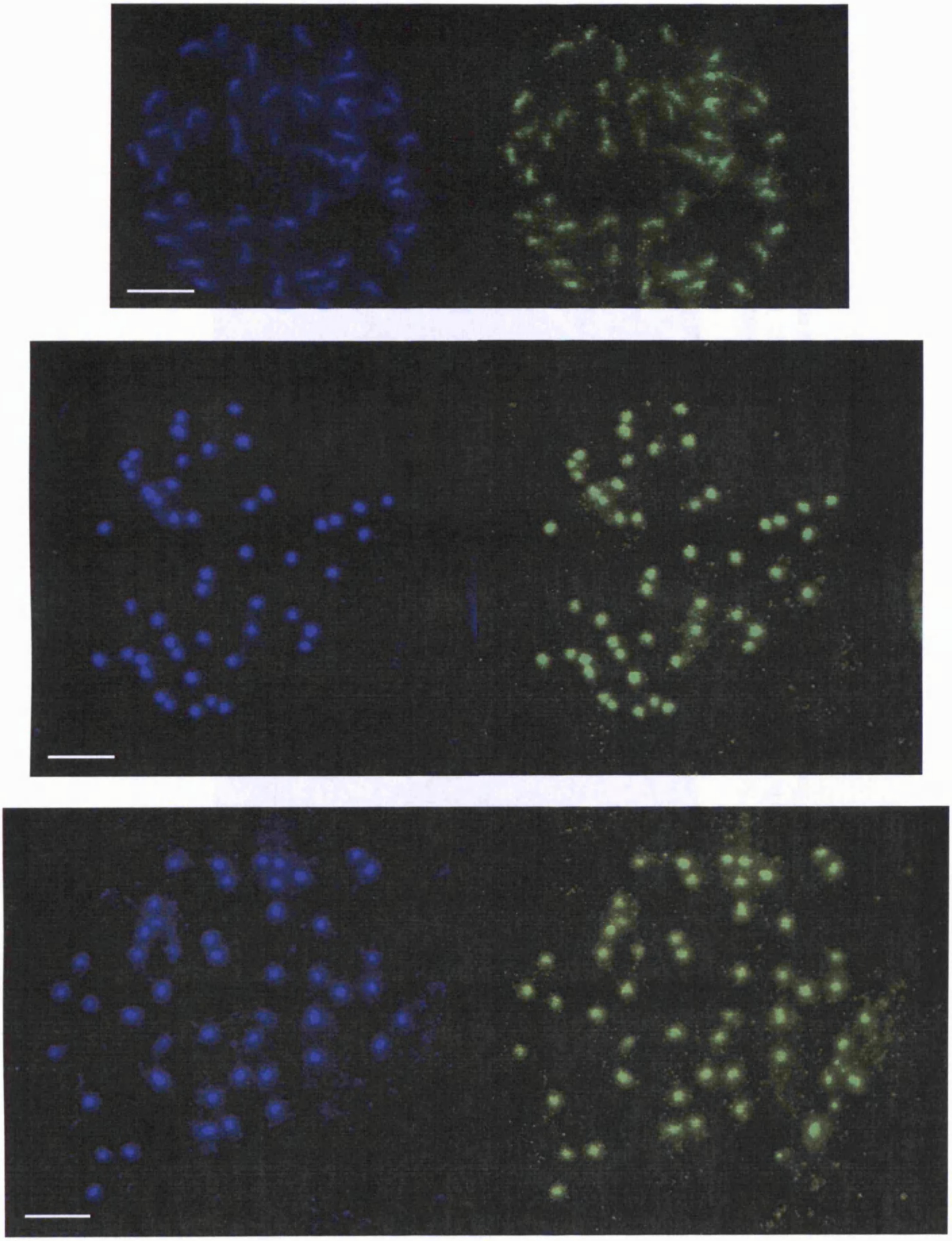


Figure 3.22. Three metaphase cells of *Solanum tuberosum* 'Desiree'. Left hand images, chromosomes stanied with DAPI. Right hand images, *in situ* hybridization with a potato specific *copia* retroelement as a PCR product labelled and detected with digoxigenin/FITC. The top pair of images has less condensed chromosomes, the others, more fully contracted chromosomes. Bars represent 10 μm.

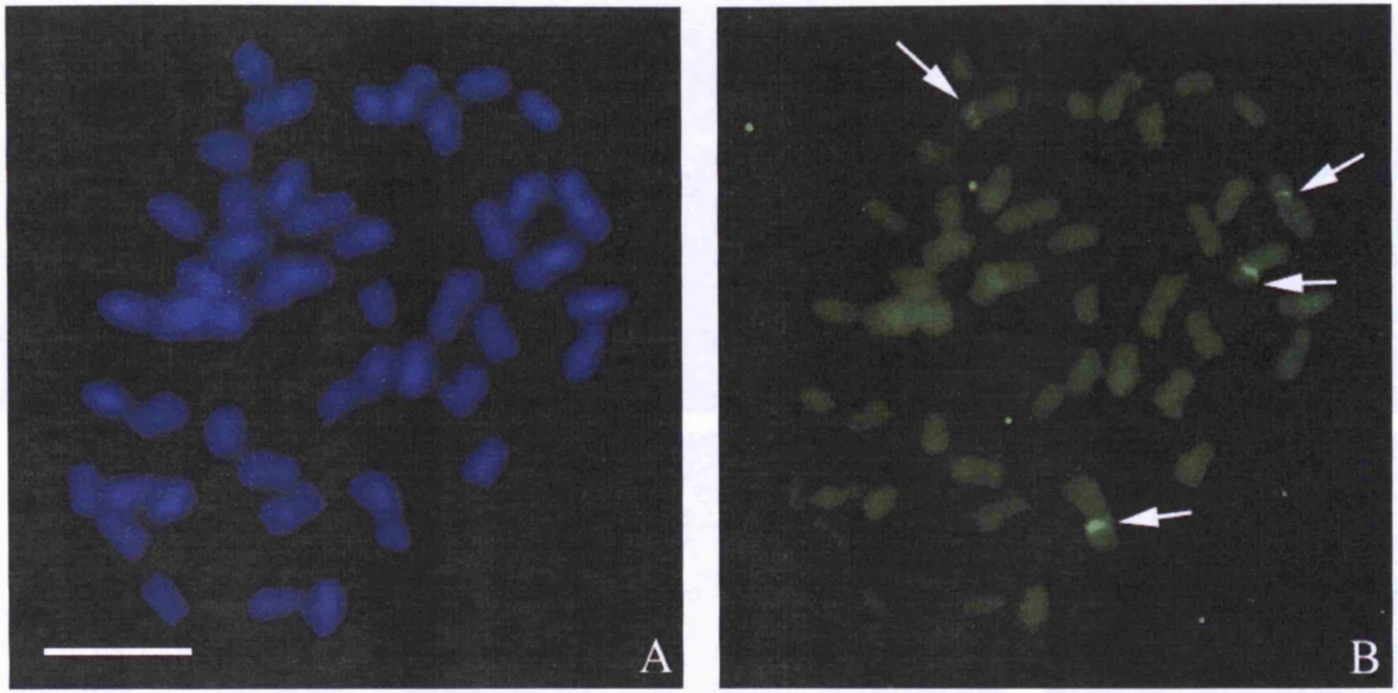


Figure 3.23. Metaphase chromosomes of *Solanum tuberosum* 'Desiree'. A) 48 DAPI stained chromosomes. B) *In situ* hybridization with a clone of 5S (pTa794) labelled and detected with digoxigenin/FITC. The localization of the four 5S signals to the short arms of four large chromosomes is marked with arrows. Bar represents 10 µm.

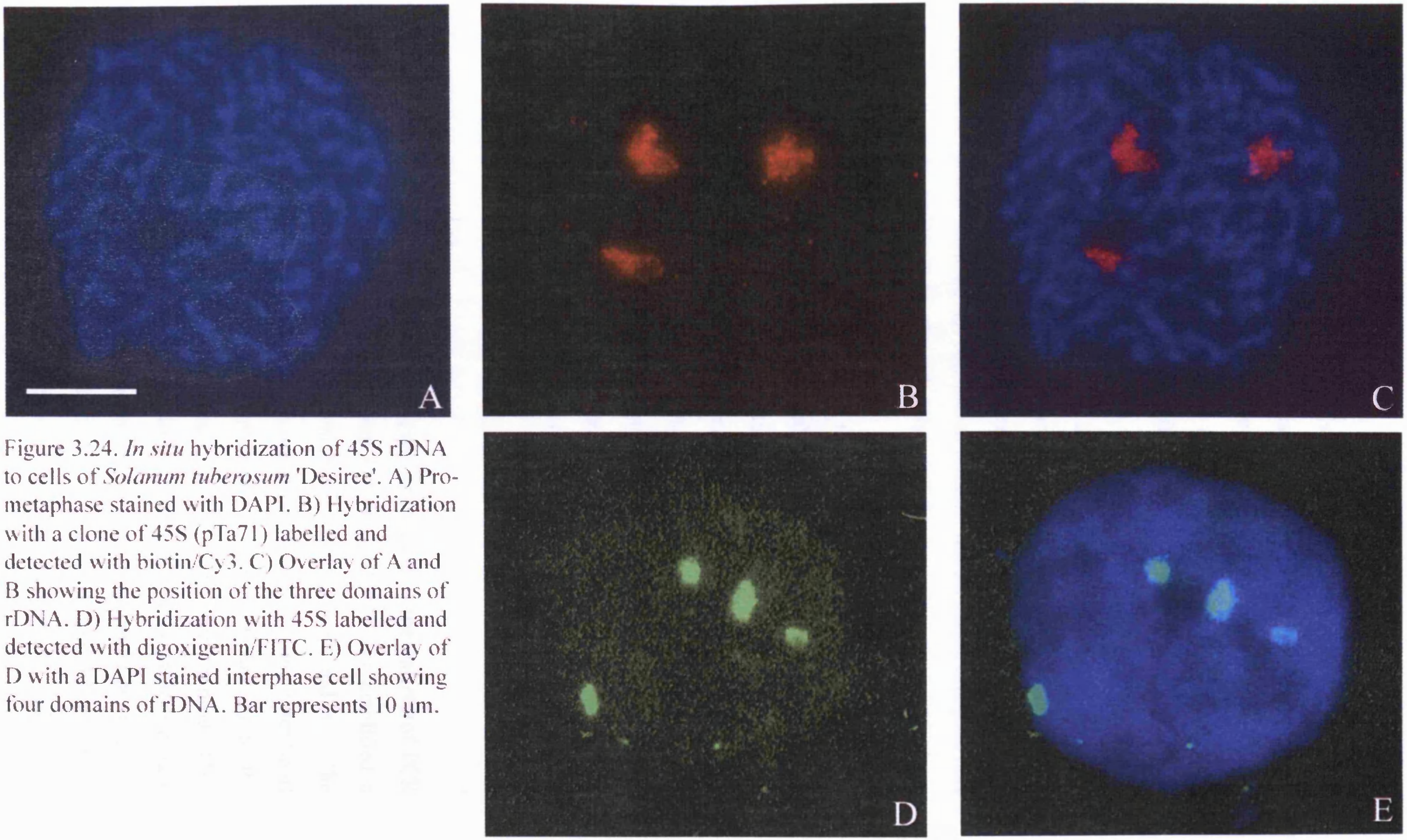


Figure 3.24. *In situ* hybridization of 45S rDNA to cells of *Solanum tuberosum* 'Desiree'. A) Prometaphase stained with DAPI. B) Hybridization with a clone of 45S (pTa71) labelled and detected with biotin/Cy3. C) Overlay of A and B showing the position of the three domains of rDNA. D) Hybridization with 45S labelled and detected with digoxigenin/FITC. E) Overlay of D with a DAPI stained interphase cell showing four domains of rDNA. Bar represents 10 μm.

3.4.1 Interpretation of phylograms

Neighbour Joining is a phylogenetic distance method, which constructs a phylogram to show branching order and express distances between sequences as varying branch length. The branch length is derived from the number of nucleic acids substitutions between sequences and corresponds to the divergence between sequences (Hall, 2001). The scale bar on the phylograms (figure 3.5 and 3.7) represents a 10% divergence. This measure is reversed in the tables 3.2 and 3.3 to show similarity between sequences. The bootstrap values are a measure for reliability or the probability that some sequences are always members of the same group. The value is generated as the number of times a certain branch appear in for instance 1000 trials. Bootstrap values below 50% are not considered reliable, but all major branches here were present nearly 100% of the time showing that the groupings are very robust. Bootstrap values are only shown on major branches, as the large grouping into sequence families is the important feature in the phylogrammes. The potato PRV-L sequences cluster into three distinct groups either as full-length clones or trimmed to half size (figure 3.4, 3.5, 3.6 and 3.7). The homologies within the groups are large, from 88-100% similarity (table 3.2 and 3.3) and the overall similarity between potato PRV-L sequences ranged from 63% to 100%. The homology to other sequences is limited, with only TVCV and TPV-like sequences having high similarity and the similarity is mostly confined to the beginning and the end of the sequences (figure 3.8 A and B). Shorter fragments of *Nicotiana otophora* and *N. tomentosiformis* are most closely related to TVCV.

3.4.2 Limitation of primers

Specific primer pairs as the ones used here give limitations to the variety of PCR sequences obtained. For instance the primers (TPV6118-TPV7072) amplified a group of potato clones (Sotu6118/7072), which the other primers did not. The priming site for Sotu6118/7072 was upstream and downstream from the other SotuI clones and it did not contain the internal sequences for appropriate binding by the other primers. Some of the primers did not readily give any amplification product but there can be various reasons for a primer not to work such as PCR conditions or lack of target in the DNA. Even though specific primers are used, a change in annealing temperature affects the degree of mismatch to the DNA sequence and the amount of product obtained. The primer site of the obtained sequences mostly had a high homology to the primers used, perhaps as the annealing temperature was in the high end for the Tm of the primers. The primers have rather low Tm (table 2.2) from 48 to 58 °C, and was run with annealing temperatures from 49-56 °C. Bringing the annealing temperature down can both give amplification of related but less homologous sequences and give unwanted non-specific amplification.

3.4.3 Expression

Endogenous PRV-L sequences might be present in the genome of plants, but are they expressed? Two different potato clones were used to search Expressed Sequence Tags (EST) databases. EBI-Fasta-ESTs search with Sotu6118/7072 gave some very significant hits:

```
The best scores are:                         length      E(12579887) %identity  over nt
1.EM_EST:BE472140  potato stolon             (495) [f]   3.1e-26       84.6       279
2.EM_EST:BE472207  potato stolon             (380) [f]   1.5e-23       81.7       279
3.EM_EST:BG134760  tomato crown gall         (581) [f]   2.8e-20       78.8       274
4.EM_EST:BG133012  tomato crown gall         (494) [r]   8.9e-19       78.0       241
5.EM_EST:BI922411  tomato callus             (549) [r]   2.2e-15       79.3       208
6.EM_EST:BG133272  tomato crown gall         (212) [f]   1.7e-09       81.8       148
7.EM_EST:AW035382  tomato callus             (488) [r]   4.2e-09       80.1       146
8.EM_EST:AW033674  tomato callus             (535) [f]   6.5e-07       80.3       127
9.EM_EST:BE459070  tomato green fruit        (591) [r]   0.00024       80.2       106
```

The list tells from left to right that the EST database has been searched, the ID number of the sequence, a short description of which kind of sequence, the length of the sequence, whether the query sequence are read in the same direction as the found sequence (f) or in the reverse order (r), the E value (The E value is a statistical measure of the significance of the match and the number gives the likelihood of a similar match occurring by chance. Generally a figure of 0.01 or below is statistically very significant) and the identity over a number of bases. In sequence 1 and 2 the tissue was taken from growing stolons and early stages of tuber formation. Sequence 3, 4 and 6 were from plants inoculated with *Agrobacterium tumefaciens* where tissue from developing galls was taken. Sequence 5, 7 and 8 were callus grown from cotyledons of 7-10 days old tomato seedlings. Sequence 9 was from tissue of developing green tomato fruits. Two different regions of Sotu6118/7072 found homology to the above ESTs, one region from 0 up to 240 bp represents the end of the TAV ORF of TPV and was mostly in the reverse orientation, the other region 590 to 890 bp is a repeat region only in forward orientation.

A few significant ESTs were found with EBI-Fasta-ESTs search for SotuI-1 most of which are the same as for Sotu6118/7072.

ESTs significantly similar to SotuIII-1 were found in EBI-Fasta-ESTs database:

```
The best scores are:                        length      E(12579887) %identity  over nt
1.EM_EST:BG889385 potato dormant tuber (440) [f]  1.6e-53      89.7        439
2.EM_EST:BQ511260 potato tissue        (444) [f]  9.3e-53      89.6        445
3.EM_EST:AW616329 tomato trichomes     (480) [f]  6.2e-49      82.7        481
4.EM_EST:BQ511261 potato tissue        (436) [r]  2.4e-39      87.5        360
5.EM_EST:AW442093 tomato red fruit     (695) [r]    6e-16      85.3        191
6.EM_EST:BE343552 potato stolon        (628) [r]  4.6e-11      79.3        193
```

Sequence 1 is tissue from potato tubers stored for one month after harvest. Sequence 2 and 4 are mixed potato tissue. Sequence 3 is trichomes (hairs) from leaves of various stages of *Lycopersicon hirsutum*. Sequence 5 is *L. esculentum* red and ripe to over ripe fruit tissue. Sequence 6 is tissues from stolons and early stages of tuber formation. The region of SotuIII which found homology to these sequences was mostly the last half about 500 to 1000 bp which are the active region of RT.

Expressed PRV-L sequences are found in EST databases but significant similarity to potato PRV-L is only found among potato and tomato ESTs. It looked like SotuIII-1 had homology to ESTs from more developed tissue than Sotu6118/7072 e.g. stolon versus dormant tuber, green fruit versus red fruit, gall and callus tissue versus mixed tissue and trichomes. But from the few samples it is difficult to say if that is a general pattern or a coincidence. The ESTs were found in both sense and anti-sense orientation. One theory about conserved presumably non-essential gene regions is that sense and anti-sense transcripts form heteroduplexes possible involved in post transcriptional processes (Hughes, 2000). Inverted repeats, which are present in the newly obtained PRV-L sequences (figure 3.11 and 3.12), can also form double stranded molecules which may be involved in gene silencing (Waterhouse *et al.*, 2001).

3.4.4 Southern and *in situ* hybridization

Labelled clones from two different areas of TPV (SotuIII, the RT-region and SotuI, end of TAV and repeat region) hybridized to digested genomic DNA and metaphase chromosomes. The Southern hybridization gave strong signals to high molecular weight DNA excluding the possibility that the signal was due only to episomal forms of pararetroviruses (figure 3.13 and 3.14). Enzyme restriction bands are further discussed in chapter V "Restriction sites". That PRV-L sequences are integrated in the potato genome was also confirmed by *in situ* hybridization showing very dispersed signals on most chromosomes (figure 3.15, 3.16, 3.17, 3.18, 3.19 and 3.20). Dispersal of a PRV-L sequence was also seen in *Nicotiana edwardsonii*

(Lockhart and Schwarzacher, unpublished data) while in *Petunia* and *Musa* the sequence was confined to a few sites (Richert-Pöggeler, Schwarzacher and Harper unpublished data; Harper *et al.*, 1999)

PCR, Southern and *in situ* hybridization experiments indicate that there are many copies of PRV-L sequences present in potato but exact estimation is not possible because of sequence divergence. From the *in situ* hybridization experiments it was notable that there was a much higher copy number of *copia* and *gypsy* in potato than of PRV-L sequences. Potato specific *gypsy* and *copia* retrotransposon probes hybridized to chromosomes in the expected pattern (figure 3.19, 3.20 3.21 and 3.22) (Brandes *et al.*, 1997; Friesen *et al.*, 2001). High number of signal outside of the chromosomes might be related to transcription of the sequence (e.g. figure 3.17).

Dong *et al.* (2000) developed a set of chromosome specific DNA markers in potato with a set a BAC clones specific to each of the 12 chromosomes. They find that chromosome identification is a major challenge in plant species with small chromosomes, which usually show fewer characteristic bands using banding techniques than are critical for chromosomal identification. They mapped the 5S rDNA genes by *in situ* hybridization proximal to the centromere on the short arm of chromosome 1 and the 45S rDNA genes were mapped to the distal ends on the short arm of chromosome 2. Chromosomes were numbered in accordance with the genetic linkage group not length. Here the 5S probe hybridized to four homologous (figure 3.23 and 3.15), in agreement with Dong *et al.* (2000). The 45S rDNA was found less useful as a chromosome marker in this study as the signal was not distinctively confined to a chromosome but spread over an area between chromosomes (figure 3.18 and 3.24).

Chapter IV: The presence and diversity of pararetrovirus-like sequences in plants

4.1 Introduction

The aim of the work in this chapter was to discover how widespread PRV-L sequences were in the Solanaceae and other plants using the primers, described in chapter III, to amplify homologous fragments from genomic DNA by PCR. To survey an evolutionarily wide range of species, DNA from plant groups including moss and liverwort, horsetail and fern as well as Gymnosperms and both Dicotyledons and Monocotyledons were used. If found, it was intended to use the obtained clones and sequences to investigate their phylogenetic relationship.

4.2 Materials and methods

For plant material see table 2.1. The PCR was carried out as in chapter II, PCR. The Southern hybridization was carried out as in chapter II Southern hybridization - non radioactive. Alignment was performed using the program ClustalW at www.ebi.ac.uk, Neighbour Joining trees were generated from ClustalX and opened with TreeView (Page, 1996). In alignments, relevant parts of the database sequences TPV (NTA238747) and TVCV (AF190123) were used together with the out-group TraeI.

4.3 Results

4.3.1 PCR

PCR to amplify PRV-L sequences was carried out using primer pair TPV6241-TPV7041 or primer pair TPV6118-TPV7072, the latter often semi-nested with TPV6118-TPV7037 (table 2.2) as described in chapter III. An example of the results with the latter strategy is shown in figure 4.1 with the first and second (semi-nested) run of a PCR with genomic DNA from 12 different plant species as template. The first round gave a variety of band sizes (200bp - 2kb) or no clearly detected fragments (figure 4.1 A). The second round gave much more uniform band sizes at around 800+ bp, including those species which had no detectable product from the first PCR round (figure 4.1 B). For both PCRs, a control sample without DNA was

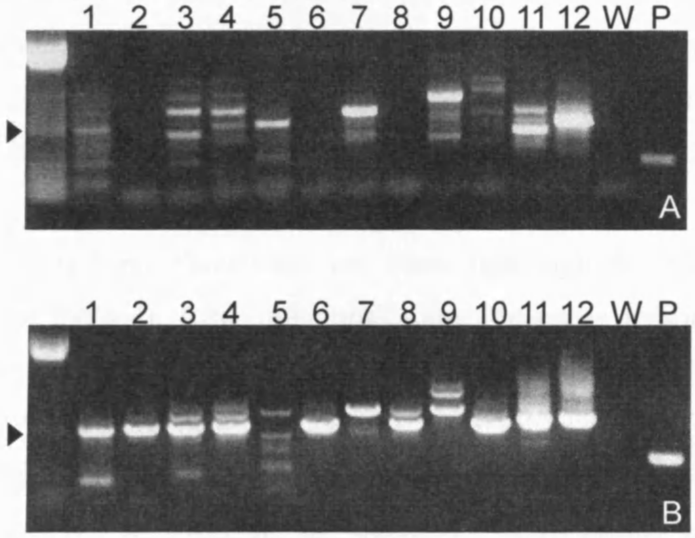

1 2 3 4 5 6 7 8 9 10 11 12 W P
A
1 2 3 4 5 6 7 8 9 10 11 12 W P
B

Figure 4.1. Size separated fragments from a PCR amplification of genomic DNA from a diverse range of plant species A) with the primer pair TPV6118-TPV7072, annealing temperature 49ºC. B) reamplification of the PCR products from A) with the primer pair TPV6118-TPV7037, annealing temperature 51ºC. In both A) and B) the arrowhead to the left of the images are at 800bp, W is the negative control without DNA, P is the positive control with 28S primers. Numbers above images correspond to the following plant species: 1. *Scleropodium purum*; 2. *Selaginella kraussiana*; 3. *Polysticum setiferum*; 4. *Gnetum gnemon*; 5. *Nuphar lutea*; 6. *Musa* 'Obino l'Ewai'; 7. *Oryza sativa*; 8. *Triticum aestivum*; 9. *Zea mays*; 10. *Pisum sativum*; 11. *Hordeum vulgare*; 12. *Solanum tuberosum*.

included. It had been necessary to include this control to ensure that the PCR products were not the result of contamination, and considerable precautions were taken to avoid transfer or aerosol transmission of DNA following preliminary experiments where the 'water' control showed an amplification product. The PCRs were generally rather problematic in terms of contamination and a lot of effort was put into trying to find a reason for it and how to control it. The precautions included working in a flow hood, cleaning equipment with ethanol and DNA-away (Fisher Scientific) and opening tubes with care. Additionally problems included some accessions which might give amplification product once, and then persistently resist further PCR. Even if some plant species did not give any amplification product (as for instance *Arabidopsis*), it cannot be ruled out that they do not contain PRV-L sequences, especially if PRV-L fragments are present in non-sequenced heteromatic regions. Possible other reasons for low or no amplification could be that the priming site showed variation or the sequence was interrupted by a stretch of other DNA.

Figure 4.2 shows results from two rounds of PCR (semi-nesting) of genomic DNA of *Scleropodium* (a moss), *Polysticum* (Soft Shield Fern), *Musa* (banana) and *Oryza* (rice) with the primer pair TPV6118-TPV7072 and reamplification with TPV6118-7037. Hybridization to Southern blots of these gels (figure 4.2 2A and 2B) with a probe from a *Polysticum* PRV-L sequence (PoseIA-8, 82% identical to TVCV, 74% to TPV) revealed that after the first round of PCR the probe only hybridized to bands from *Polysticum* and *Musa* (although the *Musa* bands were hardly visible on the gel). After re-amplification, the probe hybridized to all the strong bands except for *Scleropodium* were the hybridization seems to be to a weak band just below the stronger band.

Table 4.1 summarizes the results of the PCR amplifications from various species, and shows the PCR primers and conditions, clone names, length, nature of sequences obtained from various PCRs, and whether they are complete (5'-3') or 3'end-sequences. Fasta comparisons with the EMBL nucleotide database revealed that TPV or TVCV were the most similar sequences in the database.

4.3.2 Phylogeny

All the new obtained clones together with the potato clones described in chapter III and relevant parts of TPV and TVCV were aligned in ClustalW using default settings and trimmed to same length (figure 4.3). The forward primer TPV6241, which was

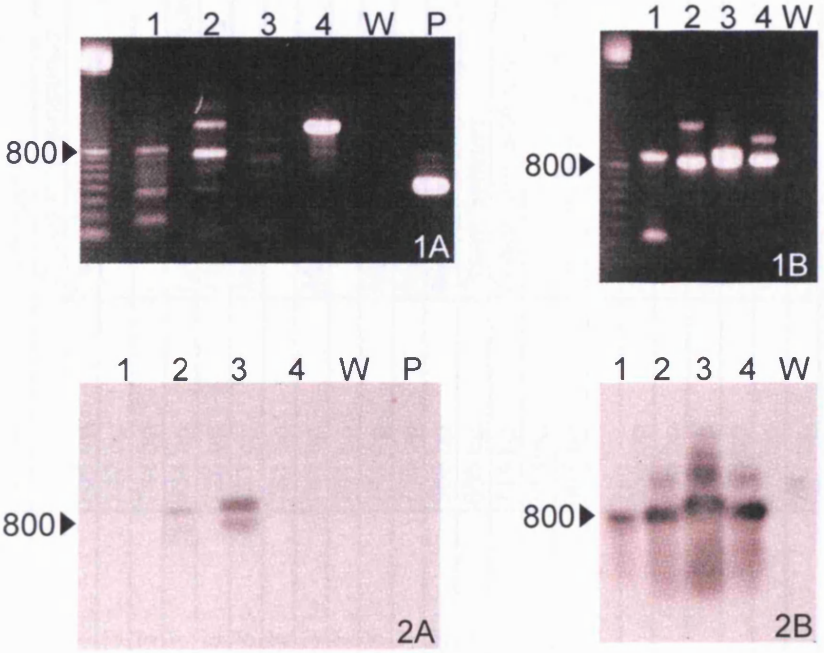


Figure 4.2. Gel images 1A and 1B) and corresponding Southern blots 2A and 2B) of two PCR runs probed with PoselA-8, a PRV-L probe from *Polysticum*. A) Genomic DNA from *Scleropodium purum* (1), *Polysticum setiferum* (2), *Musa* 'Obino l'Ewai' (3) and *Oryza sativa* (4) amplified with the primer pair TPV6118-TPV7072. B) Reamplification of the first PCR using the primer pair TPV6118-TPV7037 with the same four species numbered 1-4. W is a negative control with the same primers but without DNA, P is a positive control with primers 28S on *Oryza* DNA.

Table 4.1. Table of clones and sequences obtained from various plant species, indicating which part has been sequenced (fragment) and the length of the sequences. H_2O control, whether a negative control was included in the PCR. The forward and reverse primer used in the PCR. PCR condition gives the annealing temperature.

Clone/PCR	Fragment	Sequence length	Plant species	H2O control	F-primer	R-primer	PCR condition
Crtol-1	5' -3'	742 bp	*Crocus tommensianus*	no	TPV6241	TPV7041	56°C
Leael-1	5' -3'	734 bp	*Leucojum aestivum*	no	TPV6241	TPV7041	56°C
Leael-2	5' -3'	742 bp		no	TPV6241	TPV7041	56°C
Leael-4	5' -3'	744 bp		no	TPV6241	TPV7041	56°C
Lyes-2	5' -3'	934 bp	*Lycopersicon esculentum*	yes, neg.	TPV6118	TPV7037	49-50°C
MapolA-1	5' -3'	724 bp	*Marchantia polymorpha*	yes, neg.	TPV6241	TPV7041	?
MuBul-1	5' -3'	781 bp	*Musa* 'Butohan'	no	TPV6241	TPV7041	56°C
MuBul-2	5' -3'	781 bp		no	TPV6241	TPV7041	56°C
MuObl-1	5' -3'	741 bp	*Musa* 'Obino l'Ewai'	no	TPV6241	TPV7041	56°C
MuObl-5	5' -3'	741 bp		no	TPV6241	TPV7041	56°C
Nicll	5' -3'	741 bp	*Nicotiana clevelandii*	no	TPV6241	TPV7041	56°C
Nigll-8	5' -3'	716 bp	*Nicotiana glutinosa*	no	TPV6241	TPV7041	56°C
Nita-3	5' -3'	866 bp	*Nicotiana tabacum* 'SR1'	yes, neg.	TPV6118	TPV7037	49-50°C
PisalA-1	5' -3'	686 bp	*Pisum Sativum*	yes, neg.	TPV6241	TPV7041	?
PoselA-1	5' -3'	714 bp	*Polysticum setiferum*	yes, neg.	TPV6241	TPV7041	56°C?
PoselA-8	5' -3'	713 bp		yes, neg.	TPV6241	TPV7041	?
Trael-8	5' -3'	667 bp	*Triticum aestivum*	yes, neg.	TPV6241	TPV7041	56°C
Atropa	3'	635 bp	*Atropa bella-donna*	yes, neg.	TPV6118	TPV7037	49-50°C
Cestrum	3'	605 bp	*Cestrum aurantiacum*	yes, neg.	TPV6118	TPV7037	49-50°C
Musa	3'	683 bp	*Musa* 'Obino l'Ewai'	yes, neg.	TPV6118	TPV7037	50-52°C
Oryza-2	3'	453 bp	*Oryza sativa*	yes, neg.	TPV6118	TPV7037	50-52°C
Selaginella	3'	476 bp	*Selaginella kraussiana*	yes, neg.	TPV6118	TPV7037	50-52°C
Zea	3'	707 bp	*Zea mays*	yes, neg.	TPV6118	TPV7037	50-52°C

used to obtain most of the clones, is marked at the beginning of the alignment. Most of primer TPV7041 and all of TPV7037, as they overlap, can be seen at the end of the alignment.

The phylogram (figure 4.4) derived from the alignment in figure 4.3 was rooted with TraeI (a *gypsy*-like retrotransposon) and shows the same basic structure as when the potato clones are aligned alone (figure 3.5). The major top branch consists of all the potato clones (Sotu) divided into two subgroups and a third with Sotu6118/7072 alone. Two of these groups had very high internal homologies and were intermixed with clones from *Musa* 'Obino l'Ewai' (MuObI-1 and MuObI-5), *Nicotiana clevelandii* (NiclI-1), *Crocus tommasianus* (CrtoI-1) and *Leucojum aestivum* (LeaeI-1, LeaeI-2 and LeaeI-4). From table 3.1 and 4.1 it can be seen that the sequences in one of the groups are slightly longer than the other, about 740 bp compared to 733-735bp, except for SotuI-12 and SotuI-17. This is more evidence that they belong to different sequence families. The homology in percentage between these clones can be seen in table 4.2, each differently coloured part of the table corresponding to a group on the tree. The bottom branch of the phylogram (figure 4.4) was less homogenous and apart from TPV and TVCV consisted of clones from *Musa balbisiana* 'Butohan' (MuBuI-1 and MuBuI-2), *Nicotiana tabacum* (Nita-3), *Pisum sativum* (PisaIA-1), *Polysticum setiferum* (PoseIA-1 and PoseIA-8), *Nicotiana glutinosa* (NiglI-8), *Marchantia polymorpha* (MapoIA-1) and *Lycopersicon esculentum* (Lyes-2). The two *Musa* clones were most homologous to TPV, both about 96% and Nita-3 was 85% homologous to TPV. The closest sequences to TVCV were PoseIA-1 and PoseIA-8 with 82%. Lyes-2 was rather different to all others (table 4.2).

To include a larger sample the 3'-end sequenced PCR products were added and the alignment was trimmed to about half its length (figure 4.5). Figure 4.6 shows the corresponding phylogram. The overall structure was maintained (figure 4.4), with a homogenous group including all Sotu clones and a branch to a more diverse group including TPV and TVCV. A higher annealing temperature was used to obtain most of the clones in the Sotu group of sequences, which could be the reason for the higher homology. Together with Sotu6118/7072 there were two additional potato sequences, Desiree-1 and Desiree-2, sequences from *Selaginella kraussiana*, *Musa* 'Obino l'Ewai' and *Zea mays*. In the bottom group with TPV and TVCV, Oryza-2 from *Oryza sativa* and two Solanaceae sequences from *Cestrum aurantiacum* and

Figure 4.3. Alignment of clones obtained from potato and a variety of other plant species including relevant fragments of TPV (tobacco pararetrovirus) and TVCV (*Tobacco vein clearing virus*). TraeI a *gypsy* like retrotransposon from *Triticum aestivum* is out-group. Primer sequences are marked at the start and end of the alignment. The clustering into different groups are marked with a cross line and colours corresponding to the similarity table 4.2.

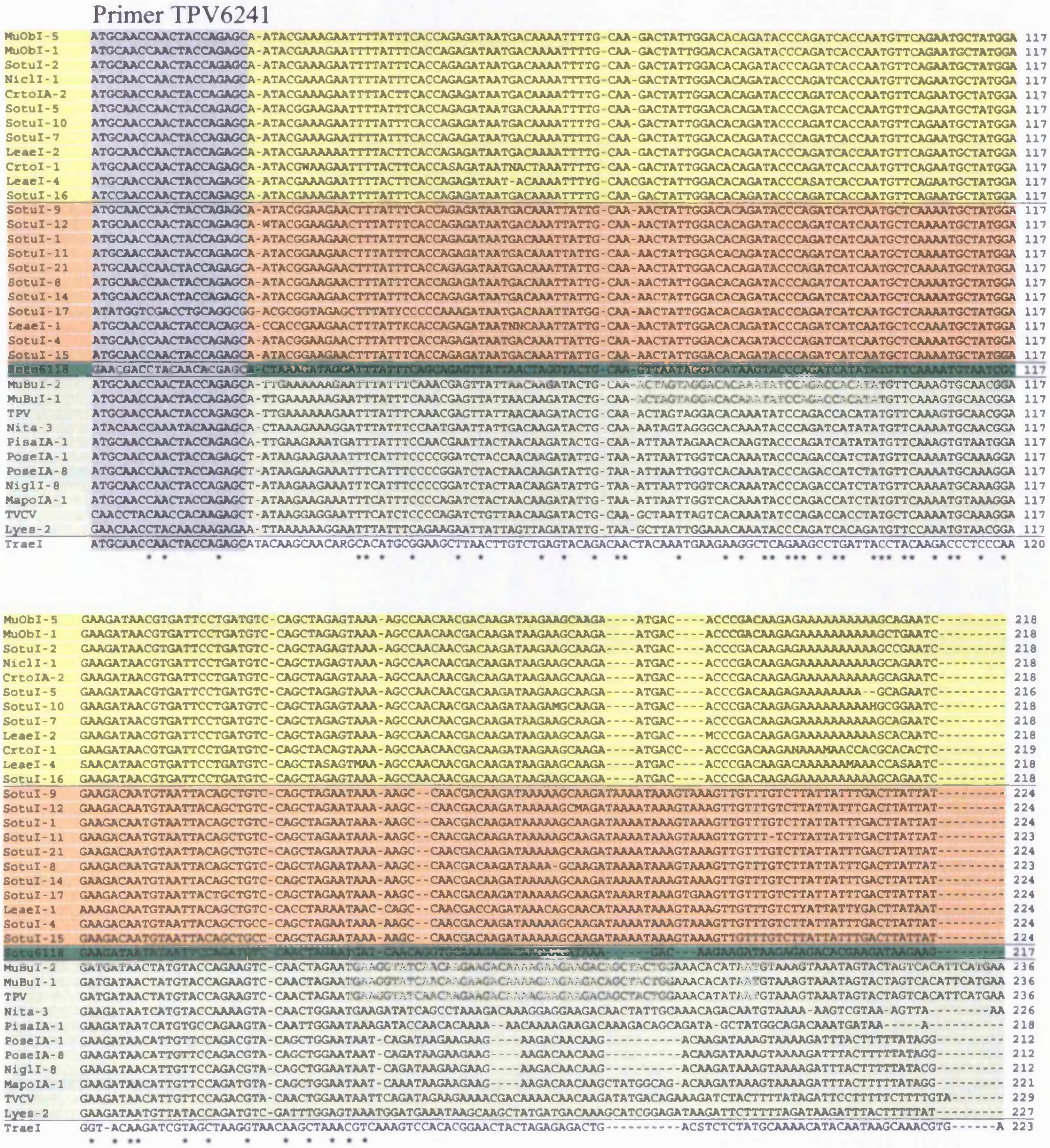


Figure 4.3. *Continued.*

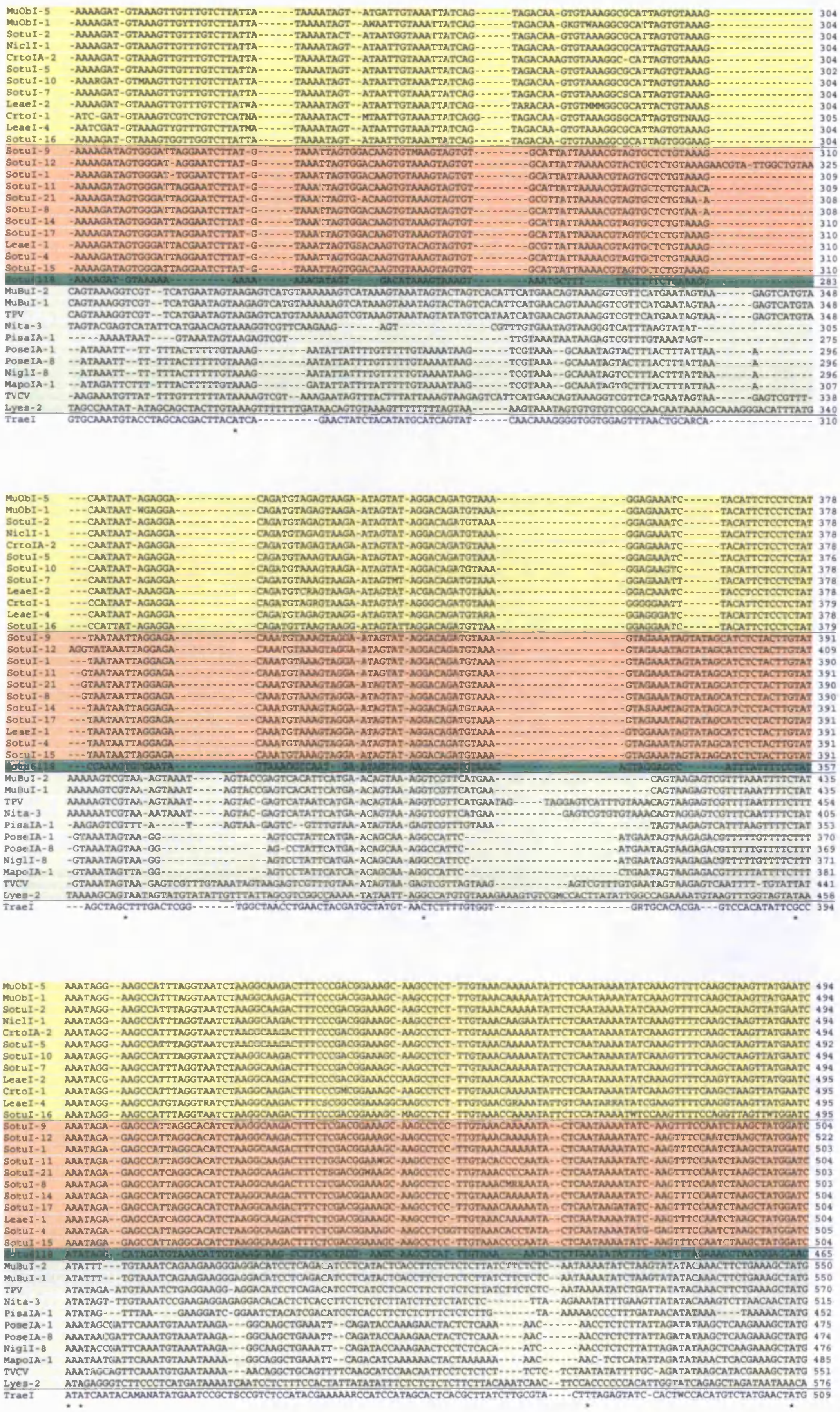
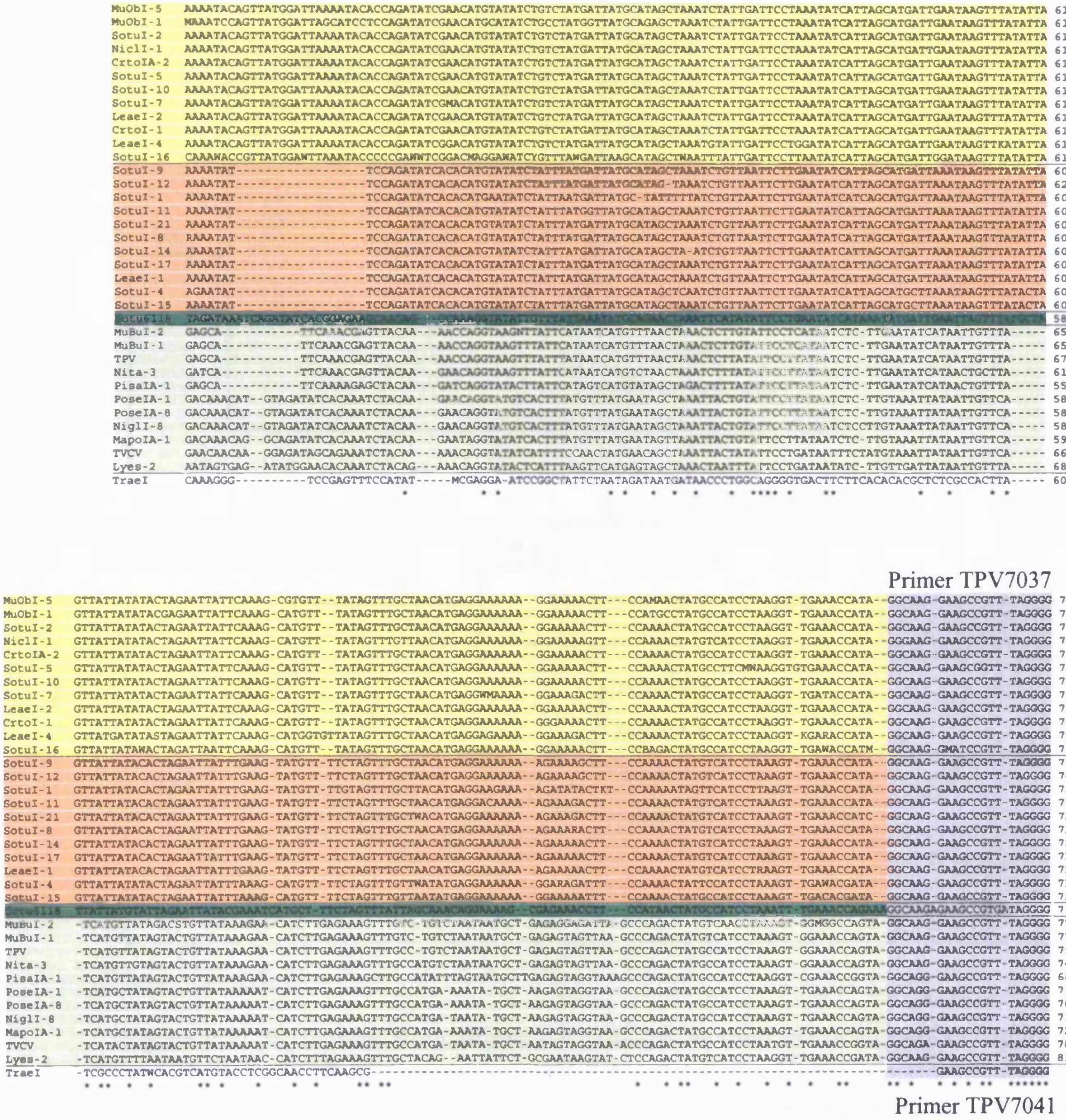


Figure 4.3. *Continued.*



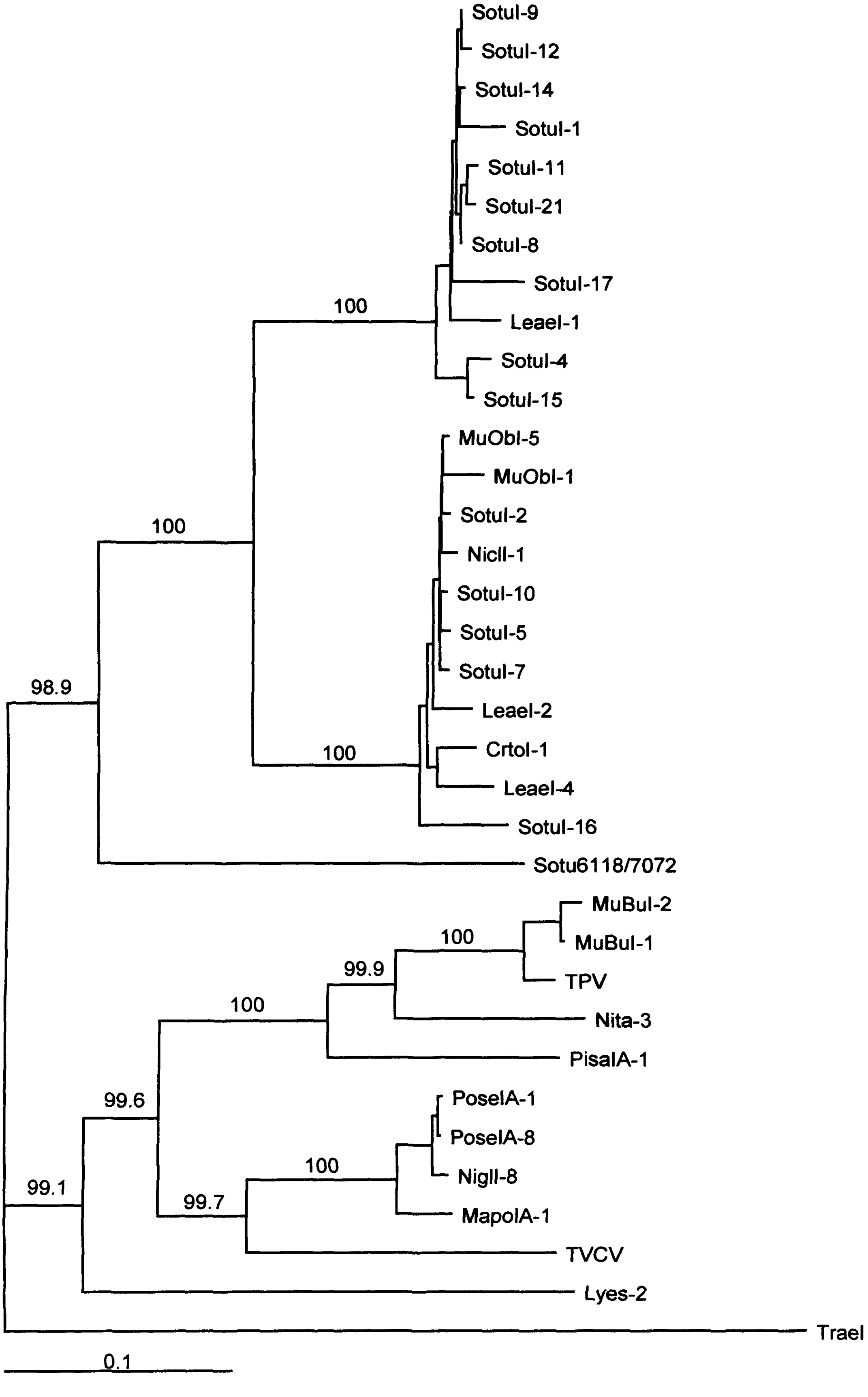


Figure 4.4. Phylogram from alignment of clones as seen in the alignment, figure 4.3. The origin of clone names are in tables 3.1 and 4.1. Numbers on major branches are the bootstrap values in percentages based on 1000 trials. Bar indicates the branch length corresponding to a 10% difference.

Table 4.2. Similarities in percentages between individual clones from the alignment (figure 4.3). The different colours represent different groups on the phylogram (figure 4.4).

	Sotul-9	Sotul-12	Sotul-14	Sotul-1	Sotul-11	Sotul-21	Sotul-8	Sotul-17	Leael-1	Sotul-4	Sotul-15	MuObl-5	MuObl-1	Sotul-2	Nicll-1	Sotul-10	Sotul-5	Sotul-7	Leael-2	Crtol-1	Leael-4	Sotul-16	Sotu6118	MuBul-2	MuBul-1	TPV	Nita-3	PisalA-1	PosalA-1	PoselA-8	Nigll-8	MapolA-1	TVCV	Lyes-2
Sotul-9																																		
Sotul-12	98																																	
Sotul-14	99	98																																
Sotul-1	97	96	96																															
Sotul-11	98	97	98	96																														
Sotul-21	98	97	98	96	98																													
Sotul-8	99	98	98	96	98	98																												
Sotul-17	94	94	94	92	93	93	94																											
Leael-1	96	95	96	94	95	95	95	92																										
Sotul-4	96	95	96	95	96	96	96	91	93																									
Sotul-15	97	96	97	95	97	97	97	93	94	98																								
MuObl-5	82	81	82	80	81	81	82	78	79	80	82																							
MuObl-1	81	79	80	79	80	79	80	76	78	79	80	97																						
Sotul-2	82	81	82	80	81	81	82	78	80	80	82	99	97																					
Nicll-1	82	81	82	80	81	81	82	78	79	80	82	98	96	98																				
Sotul-10	82	81	82	80	81	81	82	78	79	80	81	98	96	98	98																			
Sotul-5	82	81	82	80	81	81	82	78	79	80	81	98	96	98	98	98																		
Sotul-7	82	81	82	80	81	81	82	78	79	80	81	98	96	98	98	98	98																	
Leael-2	80	79	80	78	79	79	80	75	78	79	80	96	94	96	95	95	95	95																
Crtol-1	80	79	80	78	79	79	80	76	78	79	80	95	92	95	94	94	95	94	92															
Leael-4	78	77	78	77	78	78	79	74	76	77	78	93	91	93	93	93	93	93	91	91														
Sotul-16	77	76	77	75	76	76	77	73	74	76	77	92	90	92	92	92	92	92	89	89	88													
Sotu6118	65	66	65	64	64	64	65	62	64	65	65	68	68	69	68	69	68	67	68	68	65	66												
MuBul-2	65	64	65	64	64	64	64	60	63	64	65	64	64	64	65	64	64	64	63	64	62	62	32											
MuBul-1	65	64	65	64	65	64	64	61	63	64	66	65	64	65	65	65	65	65	63	64	63	63	32	98										
TPV	66	65	65	64	65	65	65	61	65	65	66	65	64	64	64	64	64	64	64	64	63	62	67	95	96									
Nita-3	62	61	62	62	62	61	61	57	61	62	63	61	60	61	61	61	61	61	59	60	59	59	66	82	83	85								
PisalA-1	65	66	65	64	65	65	64	60	62	63	66	65	64	64	64	64	63	64	64	62	64	64	68	80	81	83	82							
PosalA-1	66	66	66	65	66	66	66	62	65	66	66	65	64	65	65	65	65	65	65	64	64	63	64	71	72	74	71	73						
PoselA-8	66	66	66	65	66	66	66	62	65	66	66	65	64	65	65	65	65	65	65	64	64	63	64	71	73	74	71	73	99					
Nigll-8	66	66	66	64	67	66	66	62	64	66	66	66	65	66	66	66	65	66	66	65	64	63	64	71	72	74	72	73	98	99				
MapolA-1	66	65	65	65	66	65	65	61	64	64	66	65	63	65	65	65	65	65	65	64	64	62	65	72	73	75	70	74	95	95	94			
TVCV	64	63	64	63	64	64	64	61	62	63	64	62	61	62	62	62	63	62	61	61	61	61	66	68	70	72	71	76	82	82	81	80		
Lyes-2	63	61	63	30	63	63	62	59	61	63	63	65	64	65	65	65	65	64	64	64	62	63	20	64	65	65	64	71	70	70	71	70	67	

Figure 4.5. Alignment of clones and sequences obtained from potato and a variety of other plant species including relevant fragments of TPV (tobacco pararetrovirus) and TVCV (*Tobacco vein clearing virus*). TraeI a *gypsy* like retrotransposon from *Triticum aestivum* is out-group. The clustering into different groups are marked with a cross line and colours corresponding to the similarity table 4.3.

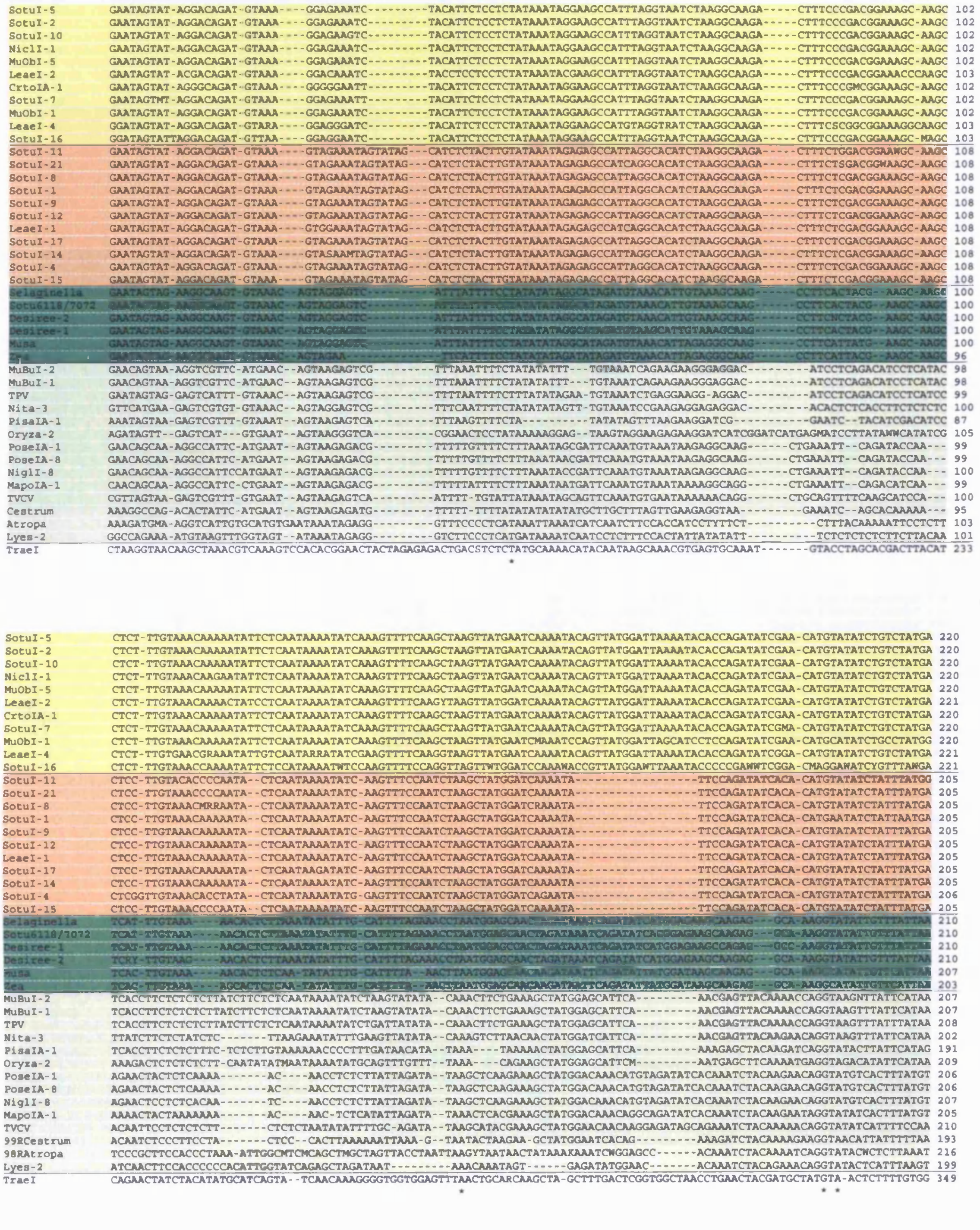


Figure 4.5. *Continued.*

```
SotuI-5        TTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAA-GCATGTT--T-----ATAGTTTGCTAACAT 332
SotuI-2        TTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAA-GCATGTT--T-----ATAGTTTGCTAACAT 332
SotuI-10       TTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAA-GCATGTT--T-----ATAGTTTGCTAACAT 332
NiclI-1        TTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAA-GCATGTT--T-----ATAGTTTGTTAACAT 332
MuObI-5        TTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAA-GCGTGTT--T-----ATAGTTTGCTAACAT 332
LeaeI-2        TTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAA-GCATGTT--T-----ATAGTTTGCTAACAT 333
CrtoIA-1       TTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAA-GCATGTT--T-----ATAGTTTGCTAACAT 332
SotuI-7        TTATGCATAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACTAGAATTATTCAAA-GCATGTT--T-----ATAGTTTGCTAACAT 332
MuObI-1        TTATGCAGAGCTAAATCTATTGATTCCTAAATATCATTAGCATGATTGAATAAGTTTATATTAGTTATTATATACGAGAATTATTCAAA-GCATGTT--T-----ATAGTTTGCTAACAT 332
LeaeI-4        TTATGCATAGCTAAATGTATTGATTCCTGGATATCATTAGCATGATTGAATAAGTTKATATTAGTTATGATATASTAGAATTATTCAAA-GCATGGTGTT-----ATAGTTTGCTAACAT 335
SotuI-16       TTAAGCATAGCTWAATTTATTGATTCCTTAATATCATTAGCATGATTGGATAAGTTTATATTAGTTATTATAWACTAGATTAATTCAAA-GCATGTT--T-----ATAGTTTGCTAACAT 333
SotuI-11       TTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAA-GTATGTT-TT-----CTAGTTTGCTAACAT 318
SotuI-21       TTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAA-GTATGTT-TT-----CTAGTTTGCTWACAT 318
SotuI-8        TTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAA-GTATGTT-TT-----CTAGTTTGCTAACAT 318
SotuI-1        TTATGC-TATTTTTATCTGTTAATTCTTGAATATCATCAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAA-GTATGTT-TT-----GTAGTTTGCTTACAT 317
SotuI-9        TTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAA-GTATGTT-TT-----CTAGTTTGCTAACAT 318
SotuI-12       TTATGCATAG-TAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAA-GTATGTT-TT-----CTAGTTTGCTAACAT 317
LeaeI-1        TTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAA-GTATGTT-TT-----CTAGTTTGCTAACAT 318
SotuI-17       TTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAA-GTATGTT-TT-----CTAGTTTGCTAACAT 318
SotuI-14       TTATGCATAGCTAA-TCTGTTAATTCTTGAATATCATTAGCATGATTAAATAAGTTTATATTAGTTATTATACACTAGAATTATTTGAA-GTATGTT-TT-----CTAGTTTGCTAACAT 317
SotuI-4        TTATGCATAGCTCAATCTGTTAATTCTTGAATATCATTAGCATGCTTAAATAAGTTTATACTAGTTATTATATACTAGAATTATTTAAA-GCATGTT-TT-----CTAGTTTGTTWATAT 319
SotuI-15       TTATGCATAGCTAAATCTGTTAATTCTTGAATATCATTAGCATGCTTAAATAAGTTTATACTAGTTATTATATACTAGAATTATTTAAA-GCATGTT-TT-----CTAGTTTGTTAATAT 318
Selaginella    ATATGCAGAACTAAATTCATATATTCCTGAATATCATAAATATGATTGAATTAGTTTATGCTAGTTATTATGTATTAGAATTATACGAAATCATGCT-TC-----CTAGTTTATTAGCAA 324
Sotu6118/7072  ATATGCAGAACTAAATTCATATATTCCTGAATATCATAAATATGATTGAATTAGTTTATGCTAGTTATTATGTATTAGAATTATACGAAATCATGCT-TT-----CTAGTTTATTAGCAA 324
Desiree-1      ATATGCAGAAGTAAATTCATATATTCCTGAATATCATAAATATGATTGAATTAGTTTATGCTAGGTATTATGGATTAGAATTATACGAAATCATGCT-TT-----CTAGTTTATTAGCAA 324
Desiree-2      ATATGCAGAAGTAAATTCATATATTCCTGAATATCATAAATATGATTGAATTAGTTTATGCTAGGTATTATGGATTAGAATTATACGRAATCATGCT-TT-----CTAGTTTATTAGCAA 324
Musa           ATATGCAAAGGTAAATTCATATATTCCTGAATATCATAGATATGATTGAATAAGTTTATGTTAGTTATTATGTATTAGAATTATACGAA-TCATGCT-TT-----CTAGTTTATTAGCAA 320
Zea            ATATGCAAAGGTAAATTCATATATTCCTGAATATCATAGATATGATTGAATAAGTTTATGTTAGTTATTATGTATTAGAATTATACGAA-TCATGCT-TT-----CTAGTTTATTAGCAA 316
MuBuI-2        TCATGTTTAACTAAACTCTTGTATTCCTCATAATCTC-TTGAATATCATAATTGTTTAT------CATGTTATAGACSTGTTATAAAGA-ACATCTTGAG-----AAAGTTTGTC-TGTC 313
MuBuI-1        TCATGTTTAACTAAACTCTTGTATTCCTCATAATCTC-TTGAATATCATAATTGTTTAT------CATGTTATAGTACTGTTATAAAGA-ACATCTTGAG-----AAAGTTTGTC-TGTC 313
TPV            TCATGTTTAACTAAACTCTTATATTCCTTATAATCTC-TTGAATATCATAATTGTTTAT------CATGTTATAGTACTGTTATAAAGA-ACATCTTGAG-----AAAGTTTGCC-TGTC 314
Nita-3         TCATGTCTAACTAAATCTTTATATTCCTTATAATCTC-TTGAATATCATAACTGCTTAT------CATGTTGTAGTACTGTTATAAAGA-ACATCTTGAG-----AAAGTTTGCCATGTC 309
PisaIA-1       TCATGTATAGCTAGACTTTTATATTCCTTATAATCTC-TTGAATATCATAACTGTTTAT------CATGTTATAGTACTGTTATAAAGA-ACATCTTGAG-----AAAGCTTGCCATATT 298
Oryza-2        TCATGCATAGCTAAACTTTTGGATTCCTTATAATCTC-TTGAATATCATAACTGTTTAT------CATGTTATAGTACTGTTATAAAGA-ACATCTTAAGT----AAGGTTTGCCATGGC 317
PoseIA-1       TTATGAATAGCTAAATTACTGTATTCCTTATAATCTC-TTGTAAATTATAATTGTTCAT------CATGCTATAGTACTGTTATAAAAA-TCATCTTGAG-----AAAGTTTGCCATGAA 313
PoseIA-8       TTATGAATAGCTAAATTACTGTATTCCTTATAATCTC-TTGTAAATTATAATTGTTCAT------CATGCTATAGTACTGTTATAAAAA-TCATCTTGAG-----AAAGTTTGCCATGAA 313
NiglI-8        TTATGAATAGCTAAATTACTGTATTCCTTATAATCTCCTTGTAAATTATAATTGTTCAT------CATGCTATAGTACTGTTATAAAAA-TCATCTTGAG-----AAAGTTTGCCATGAT 315
MapoIA-1       TTATGAATAGTTAAATTACTGTATTCCTTATAATCTC-TTGTAAATTATAATTGTTCAT------CATGCTATAGTACTGTTATAAAAA-TCATCTTGAG-----AAAGTTTGCCATGAA 312
TVCV           CTATGAACAGCTAAATTACTATATTCCTGATAATTTCTATGTAAATTATAATTGTTCAT------CATACTATAGTACTGTTATAAAAA-TCATCTTGAG-----AAAGTTTGCCATGAT 318
Cestrum        TTATGCATAGCTAAATATTTATATTCCTGATAATCTC-TTAAGTATTATGATAATT-AT------CATATTATAGTGTTGTTATAATGA-ACTGTTTGAG-----CAAGTTTGCTATAGA 299
Atropa         ACATGAGTAGCTAAATAAATTTATTCCTGAAAATCTC-TTGTTGATTWTAATTGTTTAT------CATGACTTAATAATGCTCTAATAA-GCATCTTTAG-----AAGGTTTGTTACAGA 323
Lyes-2         TCATGAGTAGCTAAACTAATTTATTCCTGATAATATC-TTGTTGATTATAATTGTTTAT------CATGTTTTAATAATGTTCTAATAA-CCATCTTTAG-----AAAGTTTGCTACAGA 306
TraeI          TGRTGCACA--CGAGTCCACATATTCGCCAT-ATCAATACAMANATATGAATCCGCTSCCGTCTCCATACGAAAAARCCATCCATAGCACTCACGCTTATCTTGCGTACTTTAGAGTATC 466
                 * *   *  *       *  **** *    **           *         * **      **     *         *    *        *           * **
```

```
SotuI-5        GAG-------GAAAAAAGGAAAAACTT----CCAAAACTATGCCTTCMWAAGGT 375
SotuI-2        GAG-------GAAAAAAGGAAAAACTT----CCAAAACTATGCCATCCTAAGGT 375
SotuI-10       GAG-------GAAAAAAGGAAAAACTT----CCAAAACTATGCCATCCTAAGGT 375
NiclI-1        GAG-------GAAAAAAGGAAAAAGTT----CCAAAACTATGTCATCCTAAGGT 375
MuObI-5        GAG-------GAAAAAAGGAAAAACTT----CCAMAACTATGCCATCCTAAGGT 375
LeaeI-2        GAG-------GAAAAAAGGAAAAACTT----CCAAAACTATGCCATCCTAAGGT 376
CrtoIA-1       GAG-------GAAAAAAGGGAAAACTT----CCAAAACTATGCCATCCTAAGGT 375
SotuI-7        GAG-------GWMAAAAGGAAAGACTT----CCAAAACTATGCCATCCTAAGGT 375
MuObI-1        GAG-------GAAAAAAGGAAAAACTT----CCATGCCTATGCCATCCTAAGGT 375
LeaeI-4        GAG-------GAGAAAAGGAAAGACTT----CCAAAACTATGCCATCCTAAGGT 378
SotuI-16       GAG-------GAAAAAAGGAAAAACTT----CCBAGACTATGCCATCCTAAGGT 376
SotuI-11       GAG-------GACAAAAAGAAAGACTT----CCAAAACTATGTCATCCTAAAGT 361
SotuI-21       GAG-------GAAAAAAAGAAAGACTT----CCAAAACTATGTCATCCTAAAGT 361
SotuI-8        GAG-------GAAAAAAAGAAARACTT----CCAAAACTATGTCATCCTAAAGT 361
SotuI-1        GAG-------GAAGAAAAGATATACTKT---CCAAAAATAGTTCATCCTTAAGT 361
SotuI-9        GAG-------GAAAAAAAGAAAAGCTT----CCAAAACTATGTCATCCTAAAGT 361
SotuI-12       GAG-------GAAAAAAAGAAAAGCTT----CCAAAACTATGTCATCCTAAAGT 360
LeaeI-1        GAG-------GAAAAAAAGAAAAACTT----CCAAAACTATGTCATCCTRAAGT 361
SotuI-17       GAG-------GAAAAAAAGAAAAACTT----CCAAAACTATGTCATCCTAAAGT 361
SotuI-14       GAG-------GAAAAAAAGAAAAACTT----CCAAAACTATGTCATCCTAAAGT 360
SotuI-4        GAG-------GAAAAAAGGARAGATTT----CCAAAACTATTCCATCCTAAAGT 362
SotuI-15       GAG-------GAAAAAAGGAAAAATTT----CCAAAACTATGCCATCCTAAAGT 361
Selaginella    ACA-------GGAAAAGCGAGAAACCTT---CCATAACTATGCCATCCTAAATT 368
Sotu6118/7072  ACA-------GGAAAAGCGAGAAACCTT---CCATAACTATGCCATCCTAAATT 368
Desiree-1      ACA-------GGAAAAGCGAGAAACCTT---CCATAACTATGCCATCCTAAAGT 368
Desiree-2      ACA-------GGAAAAGCGAGAAACCTT---CCATAACTATGCCATCCTAAAGT 368
Musa           ACA-------GGAAAAGCGAGAAACCTT---CTATAACTATGCCATCCTAAAGT 364
Zea            ACA-------GGAATAGCGAGGGGCCTT---CTATCSSTTTGCCATCCYAA[illegible] 359
MuBuI-2        TA----ATAATGCT-GAGAGGAG-ATTAG---CCCAGACTATGTCAACCTAAAGT 359
MuBuI-1        TA----ATAATGCT-GAGAGTAG-TTAAG---CCCAGACTATGTCATCCTAAAGT 359
TPV            TA----ATAATGCT-GAGAGTAG-TTAAG---CCCAGACTATGCCATCCTAAAGT 360
Nita-3         TA----ATAATGCT-GAGAGTAG-TTAAG---CCCAGACTATGCCATCCTAAAGT 355
PisaIA-1       TA----GTAATGCTTGAGAGTAGGTAAAG---CCCAGACTATGCCATCCTAAGGT 346
Oryza-2        TAAA--GTATTGCT-AAGAGTAG-TTAAG---CCCACGGTATRCCAKCCTA---- 361
PoseIA-1       AA------TATGCT-AAGAGTAG-GTAAG---CCCAGACTATGCCATCCTAAAGT 357
PoseIA-8       AA------TATGCT-AAGAGTAG-GTAAG---CCCAGACTATGCCATCCTAAAGT 357
NiglI-8        AA------TATGCT-AAGAGTAG-GTAAG---CCCAGACTATGCCATCCTAAAGT 359
MapoIA-1       AA------TATGCT-AAGAGTAG-GTAAG---CCCAGACTATGCCATCCTAAAGT 356
TVCV           AA------TATGCT-AATAGTAG-GTAAA---CCCAGACTATGCCATCCTAATGT 362
Cestrum        AAA----GTATGCT-GAGAGTAG-TTATA---TCCAGACTATGCCATCCTAAGGA 345
Atropa         AA------CATTCT-TTGAATAA-GTATG---CCCAGACTATGCCATCCTAAGGT 367
Lyes-2         AT------TATTCT-GCGAATAA-GTATC---TCCAGACTATGTCATCCTAAGGT 350
TraeI          CACTWCCACATGTCTATGAACTATGCAAAGGGTCCGAGTTTCCATATMCGAGGAA 526
                                                      *    *  *
```

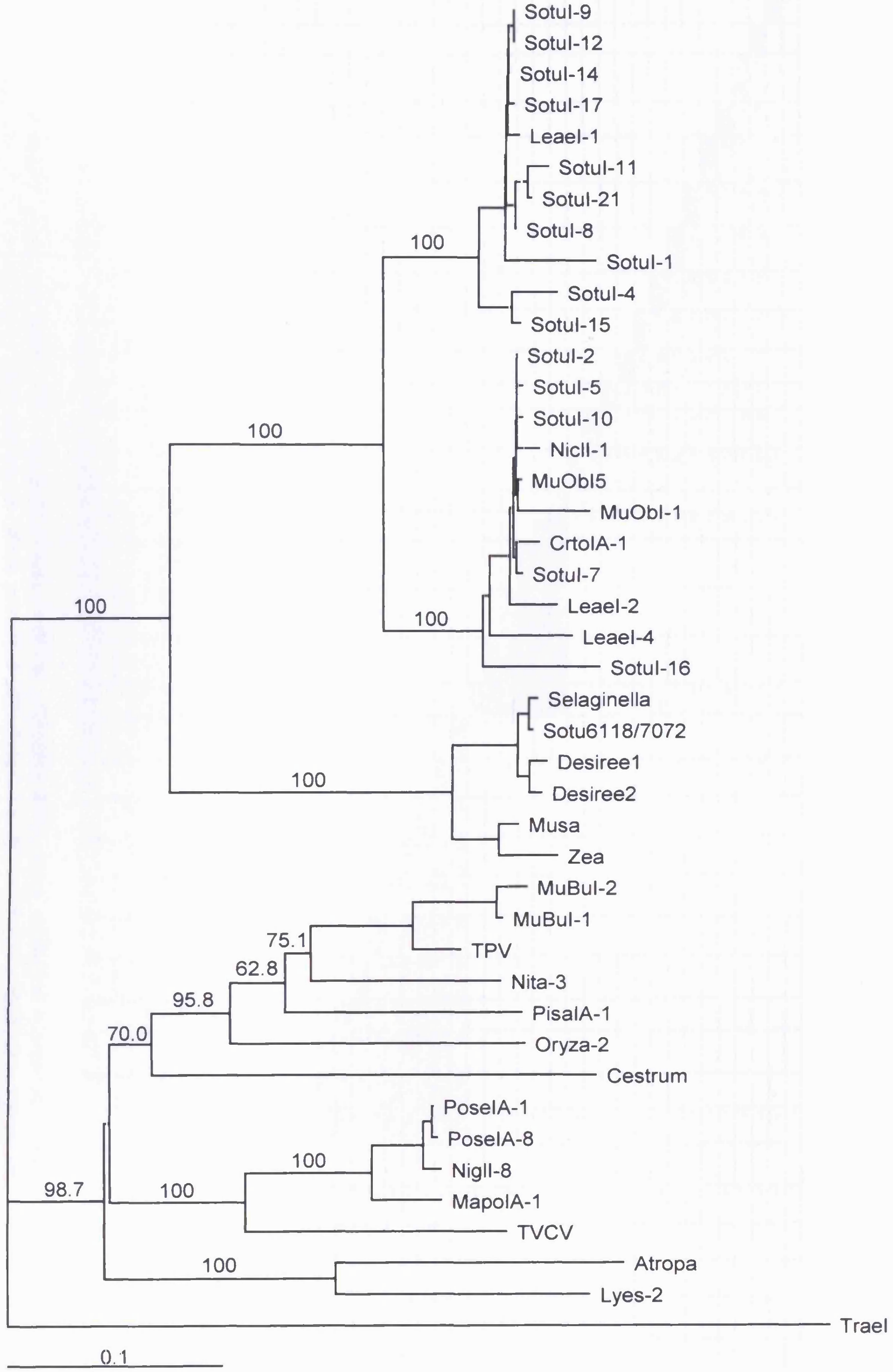


Figure 4.6. Phylogram from alignment of clones and sequences as seen in figure 4.5. The origin of clone names are in tables 3.1 and 4.1. Numbers on major branches are the bootstrap values in percentages based on 1000 trials. Bar indicates the branch length corresponding to a 10% difference.

Table 4.3. Similarities in percentages between individual clones and sequences from the alignment (figure 4.5). The different groups on the phylogram (figure 4.6) have been given different colours. Only a few of the similarities scores are given outside the coloured areas, as the tendency is the same along the rows. Numbers below *Zea* are unsure and often too low, possible because a computer error when identities are at or below 50-55%, a few hand calculated values are shown with the original low number in parenthesis.

	Sotul-9	Sotul-12	Sotul-14	Sotul-17	Lesel-1	Sotul-11	Sotul-21	Sotul-8	Sotul-1	Sotul-4	Sotul-15	Sotul-2	Sotul-5	Sotul-10	NicII-1	MuObl-5	MuObl-1	CrtolA-1	Sotul-7	Lesel-2	Lesel-4	Sotul-16	Selag.	Sotu6118	Desiree-1	Desiree-2	Musa	Zea	MuBul-2	MuBul-1	TPV	Nita-3	PisalA-1	Oryza-2	Cestrum	PoselA-1	PoselA-8	NigII-8	MapolA-1	TVCV	Atropa	Lyes-2
Sotul-9																																										
Sotul-12	100																																									
Sotul-14	99	98																																								
Sotul-17	99	99	99																																							
Lesel-1	98	99	98	98																																						
Sotul-11	97	97	96	97	96																																					
Sotul-21	97	97	97	97	97	97																																				
Sotul-8	98	99	98	98	97	97	97																																			
Sotul-1	94	94	94	94	94	92	93	93																																		
Sotul-4	93	93	93	93	93	93	93	93	90																																	
Sotul-15	96	96	95	96	95	95	95	95	91	96																																
Sotul-2	87									84																																
Sotul-5	86									83		99																														
Sotul-10	87									84		99	98																													
NicII-1	86									84		98	99	98																												
MuObl-5	86									83		99	98	99	98																											
MuObl-1	85									82		96	95	96	95	96																										
CrtolA-1	86									83		98	97	98	97	97	94																									
Sotul-7	86									84		98	97	98	97	97	94	97																								
Lesel-2	85									83		97	96	97	96	96	93	95	95																							
Lesel-4	83									81		93	92	92	92	92	89	92	92	90																						
Sotul-16	79									77		89	88	89	88	88	88	88	88	87	85																					
Selag.	68									67		70									66																					
Sotu6118	68									68		70									66		99																			
Desiree-1	67									66		68									66		97	97																		
Desiree-2	67									67		69									65		97	97	97																	
Musa	68									65		70									67		93	93	92	98																
Zea	65									64		67									66		89	89	88	88	95															
MuBul-2	55(43)									40		43									40		49				49															
MuBul-1	42									40		44									42		47				46		97													
TPV	43									42		43									41		49				48		91	93												
Nita-3	41									37		43									40		44				42		82	85	85											
PisalA-1	38									38		31									32		65				44		79	80	83	81										
Oryza-2	32									39		37									34		50(16)				12		68	70	73	56	74									
Cestrum	38									43		31									30		32				60		68	69	68	66	66	61								
PoselA-1	38									37		41									42		63				63		70	72	74	67	72	64	69							
PoselA-8	38									37		41									42		62				63		70	72	73	67	72	64	70	99						
NigII-8	38									38		41									41		63				62		69	71	74	69	69	60	69	98	98					
MapolA-1	44									37		46									40		63				64		70	72	72	67	71	63	69	94	94	93				
TVCV	38									35		40									39		63				61		67	69	72	69	74	63	68	80	80	77	78			
Atropa	37									29		41									41		22				29		33													
Lyes-2	52(36)									36		40									38		31				33		44												74	

Atropa bella-donna were added. The homology within these groups is shown in table 4.3. Only a few of the homologies between groups are given, as these numbers do not change much along a row. The numbers below the row of *Zea* are unsure and often found to be too low by comparing with the branch length on the phylogram (figure 4.6) and by hand calculating a few. This problem might be due to a computer error when identities between sequences are at or below 50-55%.

4.4 Discussion

The results presented here suggest that PRV-L sequences are, like other retroelements, a widespread and significant component of the genome of many plant species. Primers based on an endogenous TPV sequence can amplify a similar sized and homologous DNA product from a range of plants, and the specificity of amplification is increased with the use of semi-nested primers (figure 4.1 and 4.2). It is notable that specific primer sequences are able to amplify homologous sequences from a wide selection of plants, particularly because the sequence is not obviously a normal cellular gene. When a sequence is preserved like that, one might speculate that this is because of elevated fitness to the host or somehow the sequence is able to conserve itself.

Alignment (whether trimmed or not) shows that sequences from different plant families are mixed over the phylograms giving inconsistencies with known plant phylogenies (figure 4.4 and 4.6) and that some interspecies sequences are very homologous (table 4.2 and 4.3). The computer based problem of getting too low similarity values (table 4.3) might reflect a weakness in automated tools as they have difficulties detecting low homology and calculating similarities below 55%. Perhaps there is not enough overall similarity left in the trimmed alignment (figure 4.5) to detect homologies between distant groups, although the significant regions for alignment can be very short regions or dispersed nucleic or amino acids as in figure 1.5 to 1.9.

4.4.1 Phylogenetic controversies

In general terms, the dispersion of sequences from a wide range of plants is similar to that seen in comparisons of *gypsy* and *copia*-like retroelements (Flavell *et al.*, 1992a and 1992b; Friesen *et al.*, 2001). Flavell *et al.* (1992b) analysed the presence and heterogeneity of *copia* in various Solanaceous plants, especially potato, and found very different clones, up to 75% divergent in amino acid sequence within the same PCR reaction. Analysis of *copia* clones from other Solanaceae plants (*Capsicum*, *Lycopersicon*, *Petunia* and *Datura*) showed a complex intermix of species, such that some potato clones are more related to clones from another genera than to intra potato clones. This also occurs with sequences from *Pisum* and *Hordeum* being spread over the phylogenetic tree. They conclude that both vertical and horizontal transmissions must have occurred to give the presented pattern (Flavell *et al.*, 1992b). *Copia* sequences from three closely related *Vicia* species also showed a high degree of heterogeneity between sequences and *copia* elements from the different species were seen to cluster together in phylogenetic analysis. The degree of heterogeneity seemed to be correlated with element copy number (Pearce *et al.*, 1996).

Heterogeneity was also detected in this study and there are probably several different families of PRV-L sequences within a species. This was seen for potato in chapter III and in this chapter for *Nicotiana tabacum* where Nita-3 was only 85% identical to TPV. Within genera even more difference can be found as sequences from *Musa* and *Nicotiana* have representatives on several different major branches of the phylogenetic tree (figure 4.6). It is already known that *Musa* harbours several different species of another PRV related sequence, BSV (Geering, 2001; Harper *et al.*, 2002).

The phylogeny of the PRV-L sequences presented here does not follow the botanical phylogeny of plants (Qiu *et al.*, 1999; Judd *et al.*, 1999; figure 1.11). Representatives from both Monocotyledons and Dicotyledons are found together in each group and those from fern and liverwort are not found on separate branches. Such lack of correlation is in consistence with the sequences of other classes of retroelements, particularly *copia* and *gypsy* families, where there is no correlation of sequence with taxonomic position (see Friesen *et al.*, 2001), even between Gymnosperms and Angiosperms. These results have been widely taken to indicate the ancient origin and high conservation of the elements, being eventually lost or altered in some lineages (Friesen *et al.*, 2001; Stuart-Rogers and Flavell, 2001; Marín and Lloréns, 2000; Malik *et al.*, 1999) rather than horizontal transfer of sequences into evolutionary recent taxa. Friesen *et al.* (2001) suggested three models to explain retroelement distribution: I) that there was an initial burst of retroelements in a common ancestor and that these have been more or less preserved through until today; II) horizontal transfer between physically and phylogenetically distant plants; III) convergent evolution has made retroelements in different plants look alike. They thought I) was most probable because of monophyletic grouping of retroelements and III) is most unlikely because the regions of homology are long and contained in very diverse elements. II) is problematic as some evidence point to that solution but it is difficult to imagine simultaneous transfer to large numbers of species divided by large distances. Horizontal transfer has however, been suggested to explain the movement of *mariner* and *P*-elements in *Drosophila* (Daniels *et al*., 1990; Robertson, 1993) and a *copia* element between *Drosophila* species (Jordan *et al*., 1999). The horizontal transfer of endogenous retroviruses in mammals has not been confirmed, perhaps because of a well-developed immune system (see Chapter I).

4.4.2 Expression

While reports of *copia* and *gypsy* retroelements in plant genome studies are common (Kumar and Bennetzen, 1999), integrated PRV-L sequences are only described in *Nicotiana*, *Petunia*, *Lycopersicon*, *Oryza* and *Musa*. Perhaps obtained sequences are being recognized as unidentified fragments because few sequences are included in the plant databases.

As in chapter III EST databases were searched for the PRV-L sequences. Lyes-2 from tomato was used as an example as it differed from the already tested potato clones. Many of the hits were identical to the ones found with Sotu6118/7072 (chapter III) and only the few new ones will be described:

```
The best scores are:                        length    E(12579887) %identity over nt
1.EM_EST:AI486918 tomato ovary              (624) [r] 4.1e-29       73.8      347
2.EM_EST:BG134577 tomato crown gall         (428) [r] 0.0004        86.5      74
3.EM_EST:BI210191 tomato suspension culture (651) [r] 0.002         74.0      127
```

Sequence 1 is from the carpel of tomato ovary around the time of anthesis. Sequence 2 is from tomato plants inoculated with *Agrobacterium tumefaciens*, where tissue from developing galls was taken and sequence 3 was material from a tomato suspension culture. It was the same regions of Lyes-2 as for Sotu6118/7072 which was found homologous; bp 1-240 and bp 500 to nearly 1000 most in reverse orientation but also some forward.

The comparisons with EST databases revealed that some of the PRV-L sequences described here, including Lyes-2 but also NiglI-8 (*Nicotiana*), PisaIA-1

(*Pisum*), were homologous to expressed sequences obtained from tomato ovary, while Lyes-2 also showed similarity to various other tomato ESTs. Often the homology was to the reverse complement of the PRV-L sequences. It is notable that, in banana, BSV infection has been reported to occur following stress in tissue culture or low temperature (Dahal *et al*., 1998; Dallot *et al*., 2001), similar stresses to those in the dedifferentiated tissue of tomato and potato where the expressed PRV-L sequences were detected. In other plant species, many authors have reported expression of *copia* and *gypsy*-like elements in stressed tissues (Grandbastien, 1998; Hirochika *et al*., 1996), suggesting that stress may play a major role in inducing expression of virus-related sequences.

Chapter V: Genomic organization and relationships of pararetrovirus-like sequences in plants

5.1 Introduction

Chapter IV showed that using PCR, PRV-L sequences can be found in the genomic DNA of a range of plant species. This method gives only limited information about the numbers of copies of the sequences. They are restricted by the homology of the PCR primers to the target sequences and do not give information about their genomic organization. Southern and *in situ* hybridization (described in chapter III) give additional information, which can be used to understand the variation and aspects of the evolution of these sequences.

The aim of this chapter was to show the genomic organization and relationship of PRV-L sequences in a larger variety of plant families with details on Solanaceae using self and non-self probes. Southern hybridization was used to verify and extend the PCR results. This method is free of many of the artefacts and false positive results which may be found with PCR. It was also intended to use *in situ* hybridization to estimate the localization and copy number of PRV-L sequences and related retroelement families in chromosomes of different plant species.

5.2 Material and methods

The plants used, together with their nuclear DNA content (C-value) are given in table 2.1. For Southern hybridization, two different systems (non-radioactive and radioactive) were used. The radioactive hybridizations were carried out at two different stringencies, low (42°C/1xSSC/1%SDS) and middle (65°C/5xSSC/ 0.2%SDS). The stringency of the non-radioactive hybridizations was: 55°C/2M Urea/0.5M NaCl. Detailed protocols are described in chapter II. *In situ* hybridization was performed as described in chapter II on metaphase preparations from roots of tobacco, *Nicotiana tabacum* 'SR1' and tomato, *Lycopersicon esculentum*. Tobacco and tomato specific *copia* retroelements were amplified with degenerate primers, Ty-1 and Ty-2 (see table 2.2 of primer sequences) and fragments of the expected size (about 300bp) were directly labelled by PCR. Cloned PRV-L sequences, Nita-3 and Lyes-2, were labelled by nick translation and random primer labelling. Hybridization stringency was about 70%.

5.3 Results

5.3.1 Southern hybridization

The Sotu PRV-L fragments hybridized to genomic DNA digests from a wide range of plant species, showing clearly that homologous sequences are abundant in many plants (Figure 5.1, 5.2 and 5.3).

Figure 5.1A shows the ethidium bromide stained gel of DNA digested with *Hae*III and *Hin*dIII. The DNA digests where loaded to the closest approximation possible to equal genome equivalents, with the loading for *Arabidopsis* being taken as 1, and the other scaled appropriately according to their C-value (see table 2.2; e.g. 4 times more *Musa* DNA was loaded than *Arabidopsis*). For the species with large genomes where this would have further overloaded the gel and for potato giving too strong a signal it was necessary to reduce the DNA amount: From *Equisetum* the load was about 20% of the equivalent, *Ginkgo* 13%, *Pinus* 2%, *Nerine* 20%, *Triticum* 84%, *Pisum* 50% and potato 42%. After hybridization with ^{32}P labelled SotuI-1, from potato under low stringency (figure 5.1 B), a 24 hour exposure showed hybridization in *Leptobryum* (top of lanes, 2), *Equisetum* (top of lanes, 4), *Polysticum* (whole length of DNA smear, 5), *Ginkgo* (whole length, 6), *Pinus* (whole length, 7), *Gnetum* (whole length, 8), *Nerine* (whole length, 10), *Triticum* (whole length, 11), *Musa* 'Obino l'Ewai' (whole length, 12), *Pisum* (whole length, 14) and potato (whole length, also for *Hae*III, 15). *Selaginalla* (3) and *Nuphar* (9) gave very weak hybridization. A longer exposure revealed hybridization to *Marchantia* and *Arabidopsis*, so all species showed some homology to the probe. Hybridization was in general a smear, but it was notable that the hybridization occurred to larger fragments than the DNA fragment profile on the gel image, particularly in the *Hin*dIII digests. Few bands were seen after hybridization. As in chapter III, several bands were found in the potato lanes with the clearest at around 1.5kb and 2.5kb (lanes 15, not visible in this exposure).

Figure 5.2 shows the gel and hybridization images from *Hae*III digests of 9 species with 2 and 14 days exposure, with 10-fold lower loadings of the species of origin of the probe. The blot was hybridized, under low stringency, with the potato clone SotuI-2 labelled with ^{32}P. In the shorter exposure, *Polysticum* (5) and *Musa* 'Obino l'Ewai' (8) showed very strong hybridization, while weaker hybridization was seen from *Marchantia* (1), *Selaginella* (3), *Equisetum* (4), *Nuphar* (7),

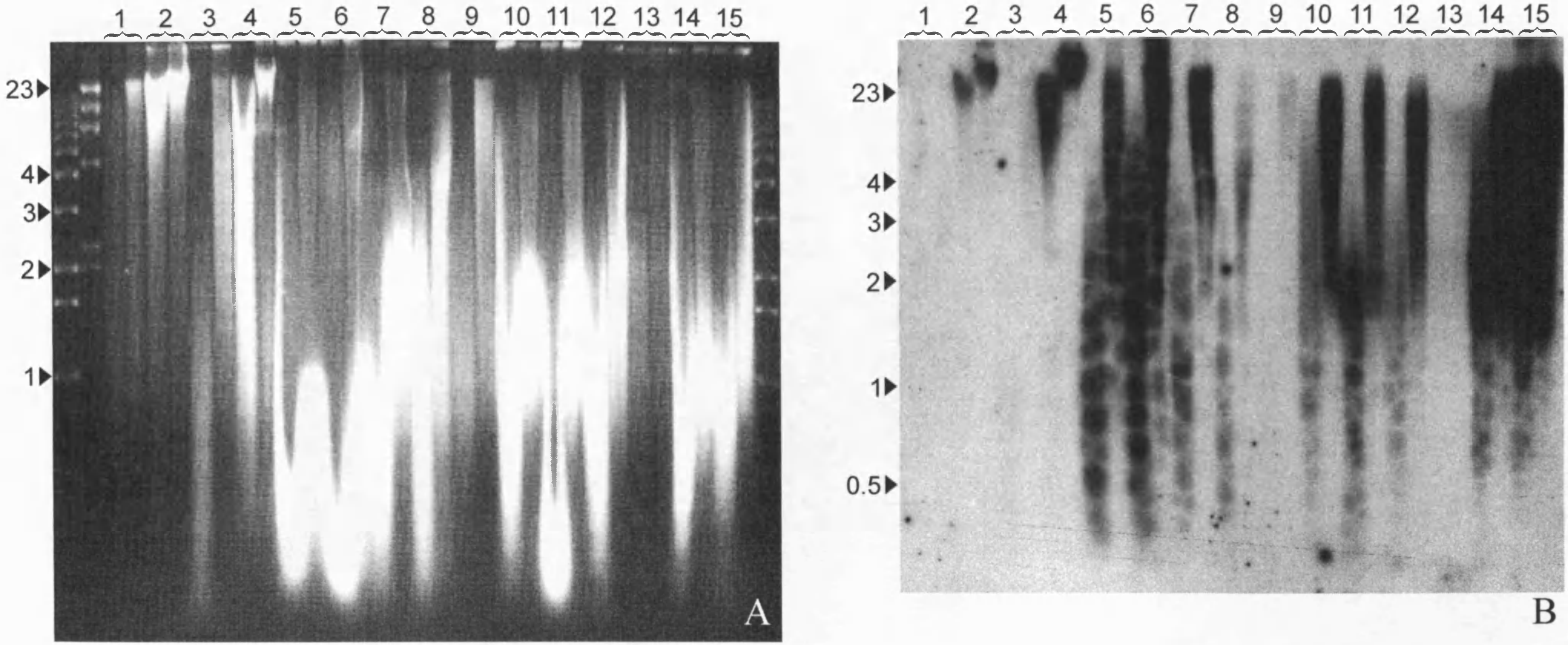


Figure 5.1. Southern hybridization with a diverse range of plant species. A) EtBr stained gel. B) Hybridization with a ^{32}P labelled clone SotuI-1. Numbers to the left of the images are size markers in kilobases, each number at the top of the images represent one plant species digested with two different enzymes, first lane *Hae*III, second lane *Hind*III. The numbers correspond to: 1. *Marchantia polymorpha*; 2. *Leptobryum pyriforme*; 3. *Selaginella kraussiana*; 4. *Equisetum arvensis*; 5. *Polysticum setiferum*; 6. *Ginkgo biloba*; 7. *Pinus pinaster*; 8. *Gnetum gnemon*; 9. *Nuphar lutea*; 10. *Nerine bowdenii*; 11. *Triticum aestivum*; 12. *Musa* 'Obino l'Ewai'; 13. *Arabidobsis thaliana*; 14. *Pisum sativum*; 15. *Solanum tuberosum*.

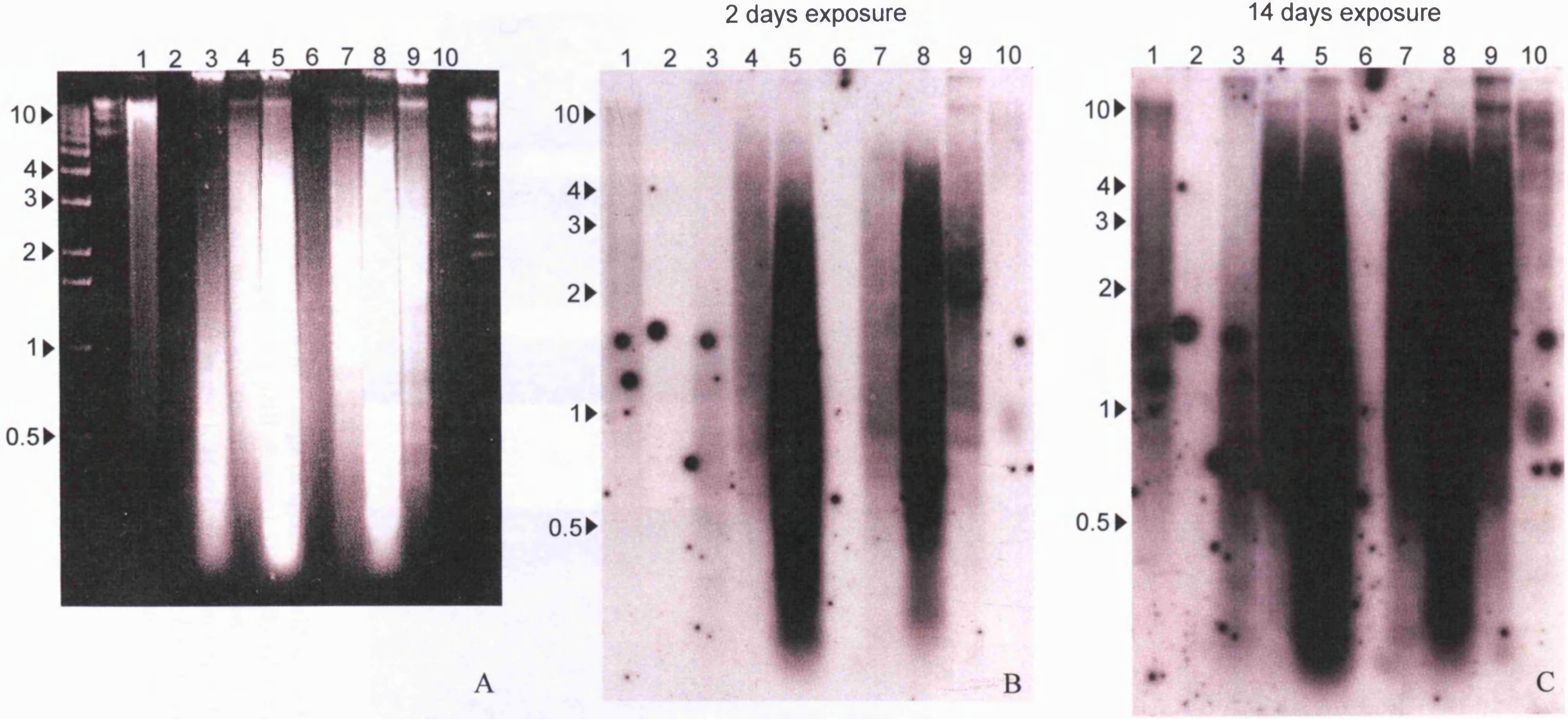


Figure 5.2. Southern hybridization to a diverse range of plant species digested with *Hae*III. A) is the EtBr stained gel of *Hae*III digested DNA. B) and C) have been hybridized with a ^{32}P labelled clone, SotuI-2 and exposed for two days and 14 days. Numbers over images correspond to the following plants: 1. *Marchantia polymorpha*; 2. Empty (*Scleropodium purum*); 3. *Selaginella kraussiana*; 4. *Equisetum arvensis*; 5. *Polysticum setiferum*; 6. *Ginkgo biloba*; 7. *Nuphar lutea*; 8. *Musa* 'Obino l'Ewai'; 9. *Arabidobsis thaliana*; 10. *Solanum tuberosum*. Numbers to the left of images are kilobase markers.

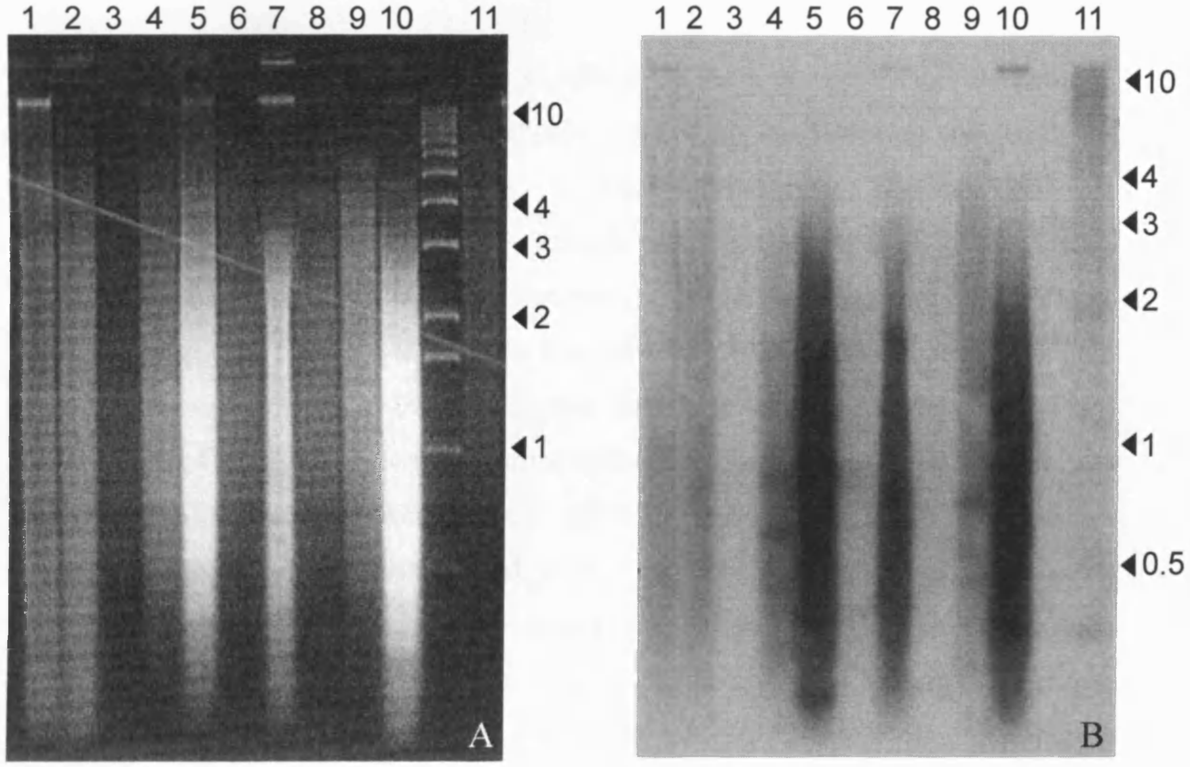


Figure 5.3. Genomic Southern blot of a range of species hybridized with a ^{32}P labelled clone, (SotuI-2). A) EtBr stained gel of *Hae*III digested DNA. B) Blot exposed for four days. Numbers to the right of images are sizemarkers in kilobases, numbers on top of images correspond to the following plant species: 1. *Marchantia polymorpha*; 2. *Scleropodium purum*; 3. Empty (*Selaginella kraussiana*); 4. *Equisetum arvensis*; 5. *Polysticum setiferum*; 6. *Ginkgo biloba*; 7. *Gnetum gnemon*; 8. *Nuphar lutea*; 9. *Musa* 'Obino l'Ewai' ; 10. *Oryza sativa*; 11. *Solanum tuberosum*.

Arabidopsis (9) and potato (10, because of lower DNA loading). After 14 days the signal was enhanced in all lanes (figure 5.2C). Too little DNA was loaded for clear detection of hybridization to *Ginkgo* (6), c.f. figure 5.1. Some bands were seen in *Musa*, smaller than 1.5 kb, and there was a band in *Arabidopsis* at 2 kb.

An additional gel and hybridization result is shown in figure 5.3 with *Hae*III digests of ten species with four days of exposure to a ^{32}P labelled clone from potato (Sotul-2) hybridized under low stringency. Strong hybridization was seen in *Polysticum* (5), *Gnetum* (7) and *Oryza* (10). Hybridization was weak to *Marchantia* (1), *Scleropodium* (2), *Equisetum* (bands are showing, 4), *Musa* 'Obino l'Ewai' (bands are showing, 9) and potato (mostly to high molecular weight DNA, and faint bands, 11).

Figure 5.4 shows the results after hybridizing at the higher stringency with a ^{32}P labelled potato clone (Sotu6118/7072). Only Potato gave a signal and that was strong to DNA above 5kb (figure 5.4 B).

To investigate the presence of PRV-L sequences within the Solanaceae, plants were selected belonging to different tribes with emphasis on species in the tribe Solaneae: *Atropa bella-donna*, *Solanum tuberosum*, *Solanum crispum*, *Lycopersicon esculentum* and *Cyphomandra crassicaulis* from tribe Solaneae; *Brugmansia* x *candida* from tribe Datureae; *Cestrum aurantiacum* and *Petunia hybrida* from tribe Cestreae; *Brunfelsia pauciflora* from tribe Salpiglossideae. Figure 5.5. 1A shows an ethidium bromide stained gel where the DNA had been digested with *Eco*RI, 1B and 1C shows the corresponding Southern blots. In 1B the blot was hybridized with AlkPhos labelled Sotu6118/7072. Strong hybridization was seen to potato (2), which had a pronounced band a bit above 3 kb. Medium to weak hybridization was seen to *Solanum crispum* (3), *Lycopersicon* (4), *Cyphomandra* (5) and *Brugmansia* (6). No hybridization was seen to *Atropa* (1), *Cestrum* (7), *Petunia* (8) and *Brunfelsia* (9). In 1C the same blot was hybridized with ^{32}P labelled Sotu6118/7072. The radioactive probe hybridized to the same species but with a stronger signal, including low molecular weight DNA and with more obvious resolution of bands. In figure 5.5 2A, 2B and 2C the DNA had been digested with *Xba*I, 2A showing the ethidium bromide stained gel and 2B and 2C the corresponding Southern blots. DNA from Petunia (8) was lost. 2B shows hybridization with AlkPhos labelled Sotu6118/7072. This probe hybridized strongly to potato (2) with two pronounced bands at about 1.5 kb and 4 kb and medium to

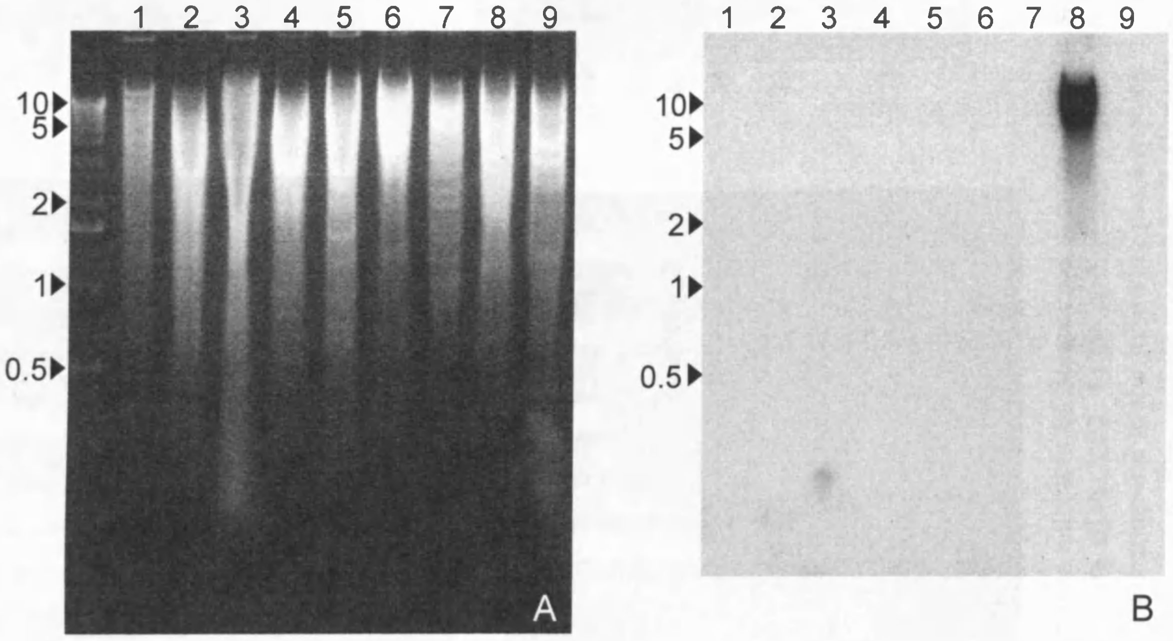


Figure 5.4. Southern hybridization with a diverse range of plant species. A) EtBr stained gel of *Hin*dIII digested DNA. B) Blot hybridized with ^{32}P labelled Sotu6118/7072. Numbers to the left of images are size markers in kilobases, the numbers at the top of images correspond to the following plant species: 1. *Selaginella kraussiana*; 2. *Polysticum setiferum*; 3. *Gnetum gnemon*; 4. *Musa* 'Obino l'Ewai'; 5. *Zea mays*; 6. *Oryza sativa*; 7. *Hordeum vulgare* 'Sultan'; 8. *Solanum tuberosum* 'Desiree'; 9. *Brassica napus*.

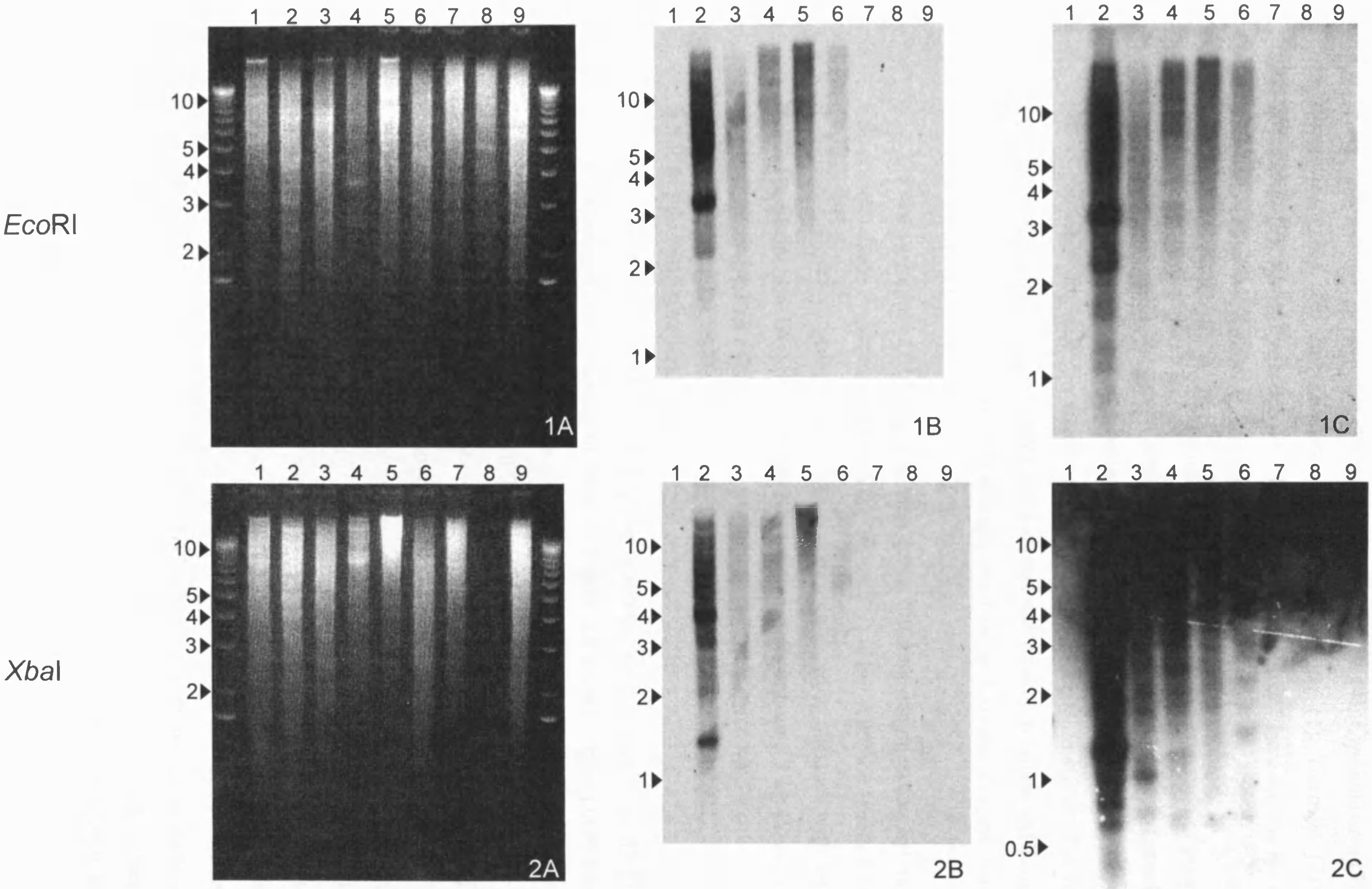


Figure 5.5. Southern hybridization to a selection of Solanaceae species. A) EtBr stained gels. B) Hybridization with an AlkPhos labelled clone Sotu6118/7072. C) Hybridization with a ^{32}P labelled Sotu 6118/7072. 1) DNA digested with *Eco*RI. 2) DNA digested with *Xba*I. Numbers to the left of images are sizemarkers in kilobases. Numbers at top of images correspond to the following plant species: 1. *Atropa bella-donna*; 2. *Solanum tuberosum*; 3. *Solanum crispum*; 4. *Lycopersicon esculentum*; 5. *Cyphomandra crassicaulis*; 6. *Brugmansia* x *candida*; 7. *Cestrum aurantiacum*; 8. *Petunia hybrida*; 9. *Brunfelsia pauciflora*.

weak hybridization to *Solanum crispum* (3), *Lycopersicon* (4), *Cyphomandra* (5) and *Brugmansia* (6). No hybridization was seen to *Atropa* (1), *Cestrum* (7) and *Brunfelsia* (9). As in figure 5.5 1C, the radioactive probe 2C gave signals to much lower molecular weight DNA and more bands. Hybridization signal strength was the same whether DNA was digested with *Eco*RI or *Xba*I, with the exception of potato.

The blot from the *Eco*RI digested DNA (figure 5.5 1 A) was re-hybridized with a recovered and purified labelled PCR product from *Cestrum* (figure 5.6). A) shows the ethidium bromide stained gel and B the hybridization signal. The probe hybridized strongly to *Cyphomandra* and *Cestrum* and weakly to potato and *Solanum crispum*. Hybridization in potato was strongly to a band at 3 kb and in *Cyphomandra* to one of about 10 kb.

Potato genomic DNA digested with seven different enzymes was hybridized with a clone from *Polysticum* (PoseIA-1) (figure 5.7) using AlkPhos labelled probe in 1B and radioactive probe in 2B. Both probes hybridized to all lanes with several bands for *Eco*RI and *Xba*I. It was especially pronounced at about 3.2 kb (*Eco*RI) and about 2.2 kb and 4.3 kb (*Xba*I)

5.3.2 *In situ* hybridization

Figure 5.8 shows hybridization of a *copia* retrotransposon in green to 48 DAPI stained metaphase chromosomes of tobacco. Figure 5.9 shows 7 blue DAPI stained chromosomes hybridized with *copia* (red). From these two figures it can be seen that the *copia* probe hybridized to most of the chromosomes, being dispersed more or less equally along the entire length of the chromosomes. In figure 5.8 about half of the chromosomes had a more intense signal than the other half.

Figure 5.10 shows tomato chromosomes stained with DAPI and hybridized with *copia* in red. In C there are 24 chromosomes, to the left lie two long chromosomes, almost separated at the NOR, a phenomenon often giving rise to a false chromosome count of 25 or 26 in tomato (26 in A). There were *copia* signals on most chromosomes, stronger and more dispersed in D, whereas in B there were fewer signals and often looking as double dots. This difference might be due to difference in hybridization stringency. In both species, *in situ* hybridization with clones of PRV-L sequences was also carried out, but no hybridization signal significantly above the background level was detected.

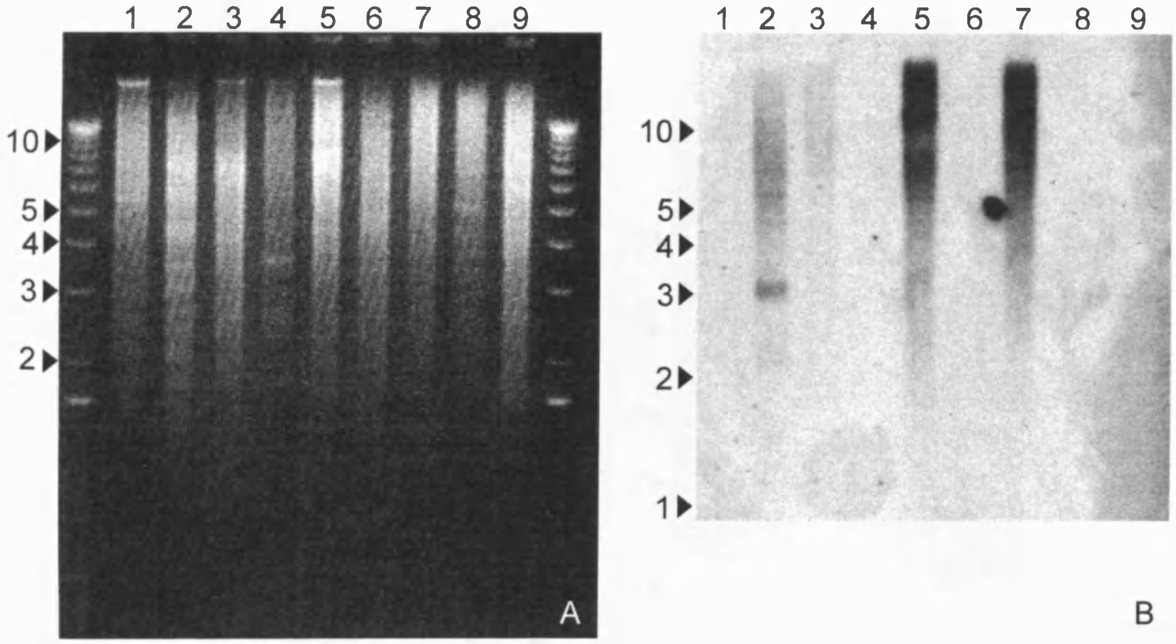


Figure 5.6. Southern hybridization to a selection of Solanaceae species. A) EtBr stained gel of *Eco*RI digested DNA. B) hybridization with an AlkPhos labelled PCR product from *Cestrum aurantiacum*. Numbers to the left of images are sizemarkers in kilobases and numbers at the top of images are the following plant species: 1. *Atropa bella-donna*; 2. *Solanum tuberosum*; 3. *Solanum crispum*; 4. *Lycopersicon esculentum*; 5. *Cyphomandra crassicaulis*; 6. *Brugmansia* x *candida*; 7. *Cestrum aurantiacum*; 8. *Petunia hybrida*; 9. *Brunfelsia pauciflora*.

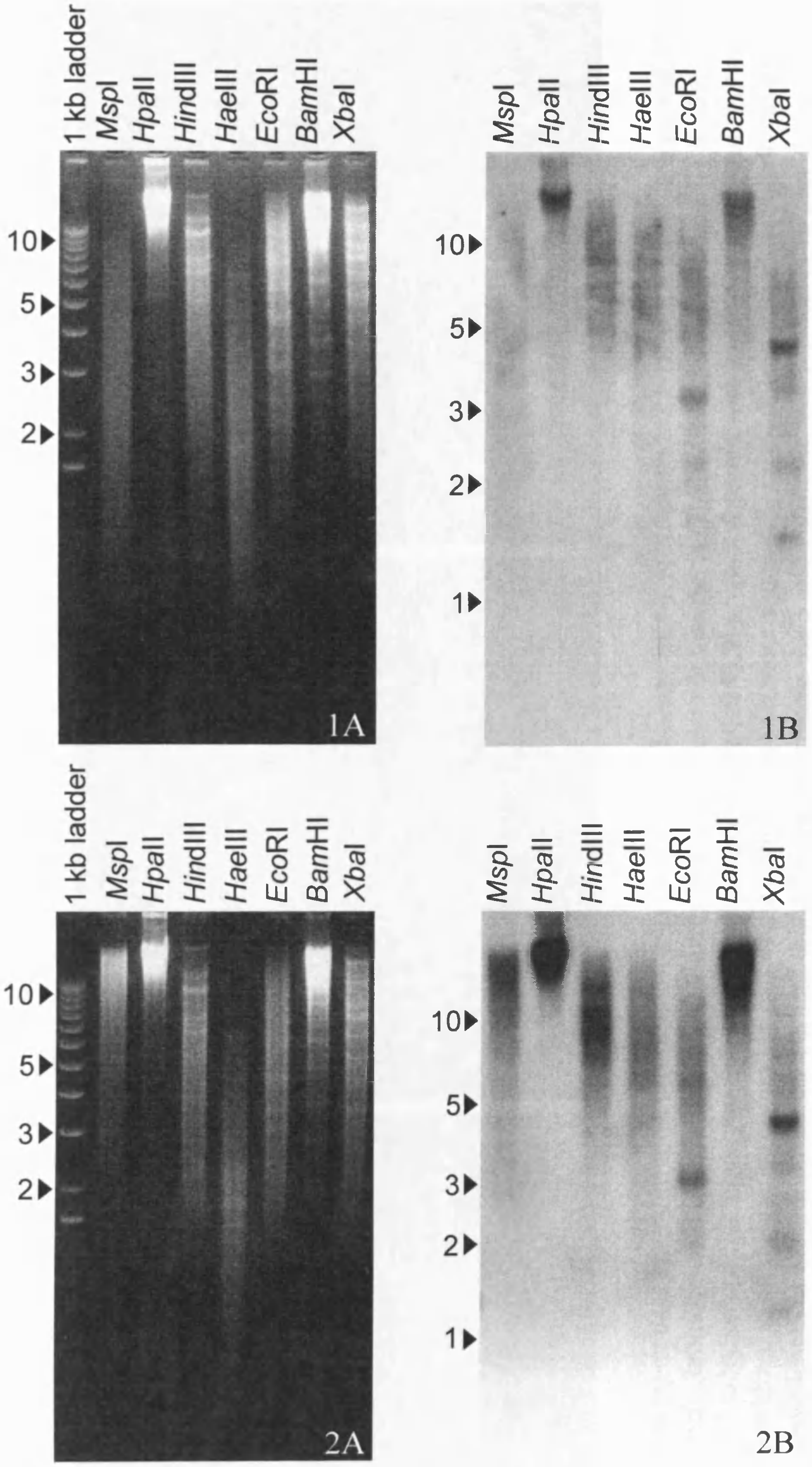


Figure 5.7. Southern hybridization to *Solanum tuberosum* genomic DNA digested with seven different enzymes. A) gel images. 1) Southern blot hybridized with AlkPhos labelled PoseIA-1. 2) Southern blot hybridized with ^{32}P labelled Pose1A-8. Numbers to the left of images are sizemarkers in kilobases.

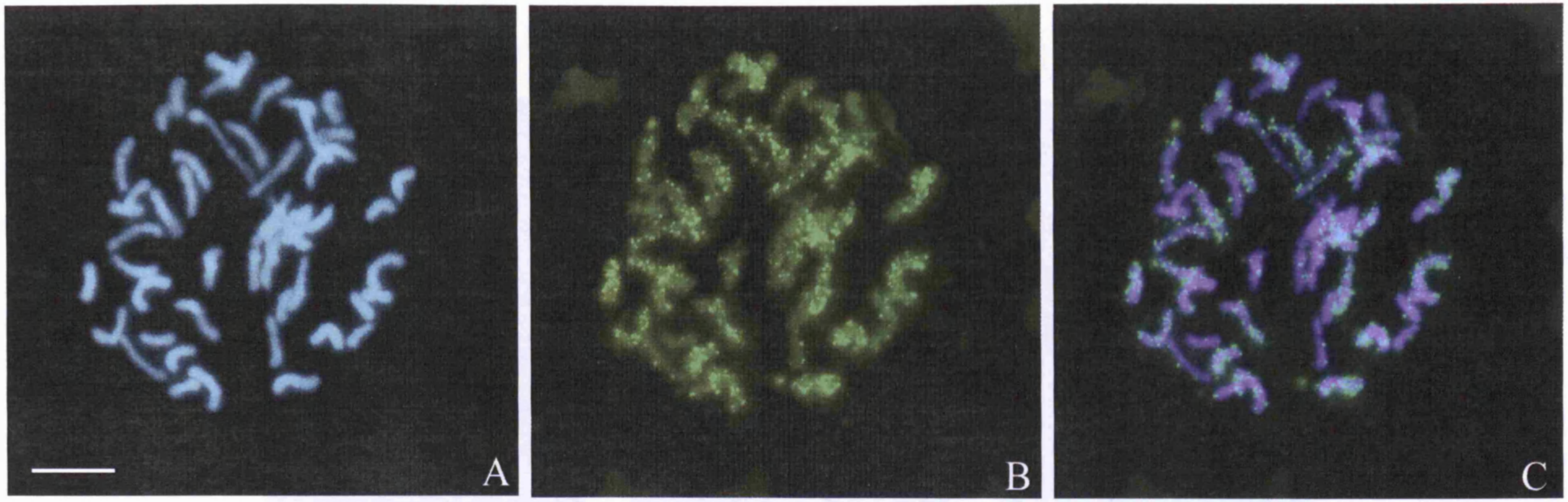


Figure 5.8. Complete metaphase spread of tobacco chromosomes (*Nicotiana tabacum* 'SR1'). *N. tabacum* has a 1C value of about 5 pg and 48 chromosomes, so each chromosome is on average some 2.5 times larger than potato or tomato. A) DAPI stained chromosomes. B) *In situ* hybridization with *N. tabacum* specific *copia* retroelement as a digoxigenin labelled PCR product detected with FITC. C) Overlay of the two. The *copia* probe hybridized to all the chromosomes but half of the complement was more abundantly labelled, and in this the signal was dispersed along the whole length of the chromosomes. Perhaps this difference represents the S and T genome. Bar represents 10 µm.

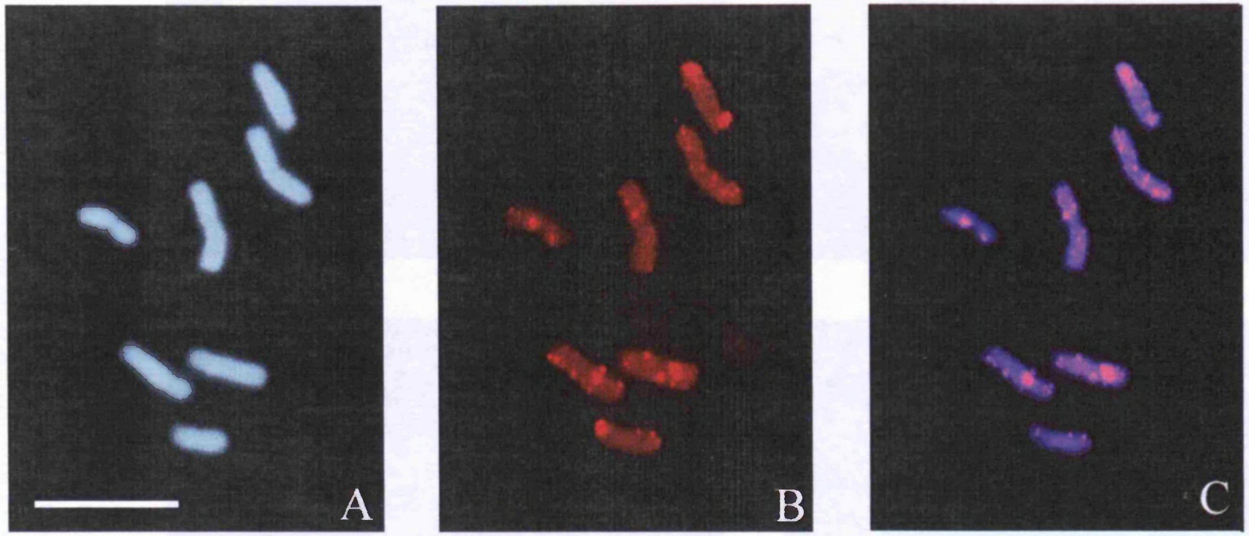


Figure 5.9. Seven chromosomes from a cell of *Nicotiana tabacum* 'SR1'. A) Stained with DAPI. B) *In situ* hybridization with *N. tabacum* specific *copia* retroelement as a PCR product labelled and detected with biotin/Alexa594. C) Overlay of A and B. The medium copy number and dispersed nature of the probe is seen. Bar represents 10 μm.

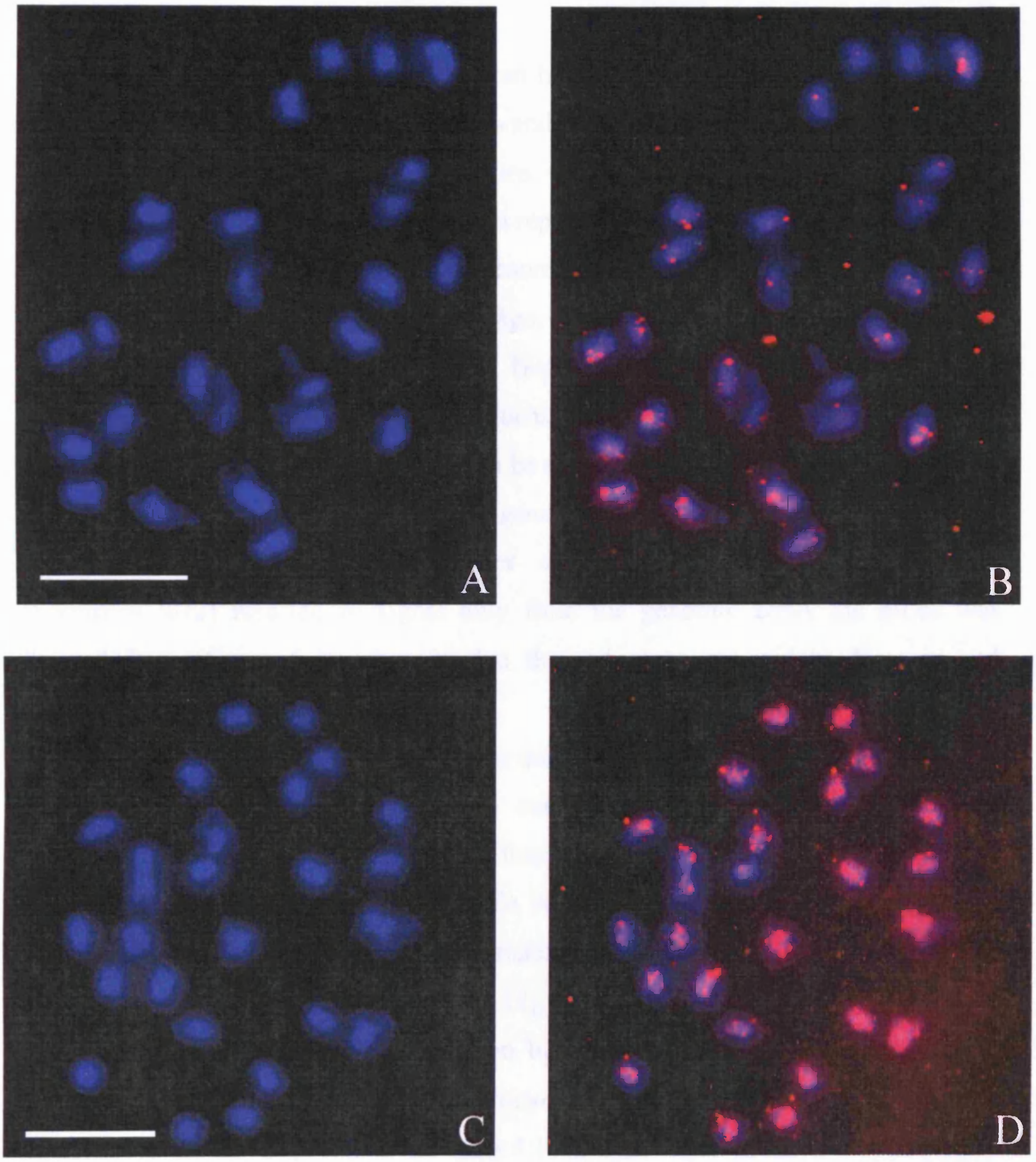


Figure 5.10. Metaphase chromosomes of tomato (*Lycopersicon esculentum*). The 1C-value of diploid tomato is 1pg with 2n=24 chromosomes, so chromosome size is similar to potato (2n=48, 1.8 pg; cf chapter I). A and B) Stained with DAPI (blue). B and D) *In situ* hybridization with a tomato specific *copia* retroelement PCR product labelled and detected with biotin/Alexa594. The hybridization shown in red, appears magenta when overlayed on the DAPI-stained chromosomes. The *copia* elements are abundant and show a dispersed distribution over most chromosomes but with reduced copy numbers in some chromosomes. Although the stringency was expected to be similar, it is evident that the upper metaphase was hybridized at higher stringency since fewer hybridization sites are seen, and they occur as double dots indicating high homology of the probe to sites on both chromatids. Bars represent 10 µm.

5.4 Discussion

5.4.1 Southern hybridization

A PRV-L probe, under low stringency, can hybridize to genomic DNA from a wide range of plant families, from lower plants and Gymnosperms to Monocotyledons and Dicotyledons. The clones used as probes were SotuI-1 or SotuI-2 from potato, homologues to ORF 4 TAV domain and a repeat region of TPV. Hybridization under low stringency conditions was seen to genomic DNA of *Marchantia*, *Scleropodium*, *Selaginella*, *Equisetum*, *Polysticum*, *Ginkgo*, *Pinus*, *Gnetum*, *Nuphar*, *Nerine*, *Musa* 'Obino l'Ewai', *Oryza*, *Triticum*, *Pisum*, *Arabidopsis* and potato (figure 5.1, 5.2 and 5.3). Therefore it looks like PRV-L sequences are a widespread constituent of the genome of these plants. Though it has to be mentioned that some of the signal form the Southern hybridization at low stringency might be the result of non-specific hybridization. Hybridization at higher stringency and with another probe (Sotu6118/7072) resulted in signal only from the genomic DNA the probe was derived from (figure 5.4), showing that the sequences are widely diverged and suggesting that a recent origin is unlikely.

Southern hybridization is not fully quantitative to allow precise estimation of copy number, but allows for general comparison. Loading of equal genome equivalents is a nice idea but is not very feasible in reality as many plants have very large genomes which means a lot of DNA is necessary and that also give problems with the limits of loading volume. For instance looking at the *Hin*dIII digest in figure 5.1 for *Nerine* (lane 10), *Triticum* (lane 11), and *Musa* (lane 12) they look to have about the same strength of hybridization but when looking at the C-values (also taking the lower load of *Nerine* and *Triticum* into account) the difference in DNA loaded from the three species are roughly 4:15:1. This implies that *Musa* and *Nerine* contain a considerable higher amount of the PRV-L sequence than *Triticum*.

Within the family Solanceae a specific PRV-L probe only hybridized to some species: the potato-derived probes analysed hybridized to genera within their own tribe (Solaneae), except *Atropa*, and to *Brugmansia* (Datureae) (figure 5.5). A *Cestrum* (Cestreae) derived PCR pool probe hybridized to its own DNA, and three genera from Solaneae; potato, *Solanum crispum* and *Cyphomandra* (figure 5.6). It is notable that a *Polysticum* derived probe hybridized to potato genomic DNA (figure 5.7) when a potato probe did not hybridize to all genera within Solanceae under the same hybridization conditions. Presumably these clones represent different families of PRV-L sequences although *Cestrum*, *Lycopersicon*, *Atropa* and two *Polysticum* clones cluster on the same major branch different from the potato branch (figure 4.6).

When Southern blots were hybridized with either AlkPhos labelled probe or radioactively labelled probe (figure 5.5 and 5.7) the overall hybridization pattern was the same, however the radioactively labelled blot can generate more intense signal and also to low molecular weight DNA.

Within the Solanaceae, the results presented here are compatible with those of Jakowitsch *et al.* (1999) who first suggested the major contribution of PRV-L sequences to the genome of *Nicotiana*. Southern hybridization with a 800 bp fragment from ORF3 in TPV to *N. tabacum* and the progenitors *N. sylvestris* and *N. otophora* or *N. tomentosiformis* showed presence of the sequence. No change in hybridization pattern was seen between *N. sylvestris* and *N. tabacum* indicating that the PRV-L sequences have remained unchanged since the formation of *N. tabacum* (Mette *et al.*, 2002). Two additional Solanaceous plants, *Datura* and *Lycopersicon*, also showed hybridization although *Petunia* (Solanaceae), *Arabidopsis* and *Pisum* were negative with the TPV probe (Jakowitsch *et al.*, 1999). With a full-length clone of TVCV, Lockhart *et al.* (2000) showed the presence of the sequence in the hybrid species *Nicotiana edwardsonii* and its male parent *N. glutinosa*; *N. tabacum* and *N. rustica* gave weak signal. No signal was detected in the female parent *N. clevelandii,* or in various other Solanaceae species including eggplant, tomato, *Petunia*, *Physalis*, and corn, soybean, wheat and turnip (Lockhart *et al.*, 2000).

5.4.2 Restriction sites

Plant pararetrovirus sequences are typically 8kb long (Hohn and Fütterer, 1997; de Kochko *et al.*, 1998). Hybridization often showed restriction bands in potato, sometimes with a length of 1.5-2.5 kb, perhaps indicating the presence of an internal fragment of the element present in multiple copies. All species showed Southern blot hybridization as a continuous smear, suggesting that multiple copies of the element were present with varying flanking sequences, at least in the presence of restriction sites for the enzymes used, with respect to the probe fragment. In Southern blot hybridization it is difficult to predict whether probes will hybridize to a specific size genomic DNA. The closest sequenced relatives to the endogenous viruses in question are the two from *Nicotiana tabacum*, NTA238747 and AF190123. A hypothetical digest of these sequences with the various enzymes used gave different results for the two because they vary in the nucleotide sequence, perhaps indicating the result will also be different for PRV-L sequences in other genera and families. These endogenous virus sequences may be incomplete in sequence and have deletions and other rearrangements. If one assumes that the virus has been inserted in its full length starting at nucleotide 1 it turn out that the domains in TPV and TVCV homologous to the probes used often are in a flanking region of unknown size after digestion. Considering the SotuI priming region (blue bar in figure 5.11) this was the case for *Eco*RI and *Xba*I for both viruses, *Xba*I is not cutting within TVCV. *Hin*dIII gave an internal fragment of 6437bp and 3086bp respectively for TPV and TVCV which are homologous to the probes used. TVCV digested with *Msp*I or *Hpa*II gave an internal fragment of 4404 bp homologous to the probes used.

Considering the SotuIII priming site (red bar in figure 5.11) an *Xba*I fragment of 3500bp will contain the probe for TPV as will a 6437bp fragment of *Hin*dIII. In TVCV a 3 kb *Hae*III fragment will contain the probe. If these sizes of bands are present after Southern hybridization they are not among the strongest, though there is a fragment of about 6 kb in the *Hin*dIII digested lane in figure 3.14 probed with SotuIII.

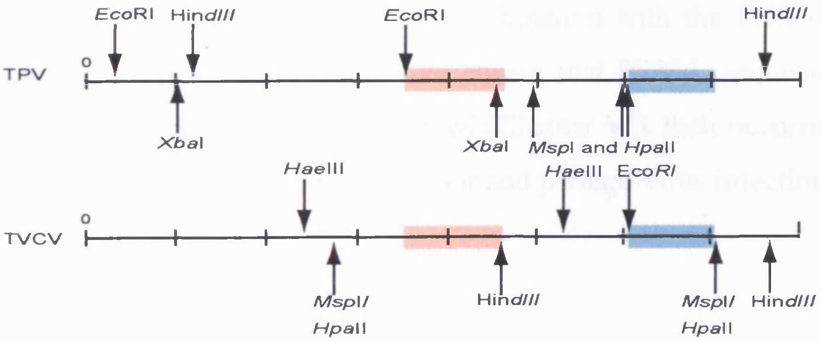


Figure 5.11. Restriction sites of TPV and TVCV both about 8 kb. The blue bar is the area homologous to the SotuI priming region, the red bar is the area homologous to the SotuIII priming region.

A comparison of the strong bands from *Eco*RI and *Xba*I digestion of potato in this chapter and chapter III reveals that SotuI-12, Sotu6118/7072, PoseIA-1 and *Cestrum* PCR product (all from the same PRV-L region) hybridize strongly to a 3 to 3.2 kb *Eco*RI fragment while SotuIII hybridize strongly to a bit lower band between

2 and 3 kb, about 2.3 kb. The strong *Xba*I band is found between 4 and 5 kb (about 4.2-4.3 kb) for all probes except maybe for Sotu 6118/7072 where it is 4 kb. This shows that the same subpopulation of *Xba*I fragments contain homologous to both the TAV and the RT region while they are contained in different subpopulations of *Eco*RI fragments.

5.4.3 *In situ* hybridization

The *in situ* hybridization results showing the dispersed locations of *copia*-like retroelements are consistent with those in other species (Brandes *et al.*, 1997), and show both their widespread genomic distribution and high copy number. Here is shown the localization of *copia* on tobacco and tomato chromosomes (figure 5.8, 5.9 and 5.10) as a medium copy number repeat. Despite positive hybridization and amplification detected by Southern and PCR, no signal was reproducibly obtained using the PRV-L probes to either tobacco or tomato chromosomes. Tobacco was not included in the Southern experiments (figure 5.5) but tomato (despite lower loading) did show a weaker signal than potato. This indicate a lower copy number of PRV-L in tomato than potato, which might be very dispersed in small fragments and the resolution of the *in situ* hybridization might not be good enough to detect a signal.

Before 1996 (Richert-Pöggeler *et al.*, 1996) there was no indication of PRV-L sequences in plant genomes. In combination with the PCR data (Chapter IV), the Southern hybridization data here shows that PRV-L sequences are widespread in plant genomes. As will be discussed (Chapter VI), their occurrence has consequences for genome evolution and behaviour and perhaps virus infection.

Chapter VI: Discussion

6.1 Techniques

A range of techniques was used in this work to characterize host DNA copies of PRV-L sequences (Chapter II, III, IV and V). These techniques are complementary, each with their advantages and disadvantages; together they have ability to obtain and characterize families of sequences, determine their distribution and physical and spatial relationship in the plant genome.

6.1.1 PCR

Since the development of the PCR (Saiki *et al.*, 1985 and 1988), it has become the method of choice for isolation of DNA sequences from genomic DNA based on the homology of the genomic sequences to short pieces of self or non self DNA. Methods of primer design and optimisation (e.g. Primer3, www-genome.wi.mit.edu/ cgi-bin/primer/primer3.cgi/primer3_www.cgi) have been improved so that efficient degenerate or specific primers can be generated from conserved regions to span sequence domains. The specificity will improve with a higher annealing temperature but a high annealing temperature is not desirable when trying to obtain less homologous sequences or when using degenerate primers.

As contamination is a considerable problem in these kinds of studies it is important to take the greatest care when approaching new PCR experiments. Being aware of the problem right from the start, which means that during plant sampling and DNA isolation, one must ensure no cross contamination occurs. Good quality and quantity of DNA is important as secondary components can inhibit enzymes. The lab distilled water may contain contaminants so a commercial source of water might be better. As for the PCRs, it could be necessary to use separate sets of equipment, UV-treat equipment, use plugged tips, change lab coat, keep pre and post PCRs separate and eventually use a another lab to set up the PCR. It is also important to dilute and aliquot the primer stock in a sterile environment and change PCR reagents regularly.

In this study, seven of the primers designed were able to amplify PRV related DNA from potato and a range of other plant species. All PCRs were run with specific primers from a PRV-L sequence in contrast to most other research with the aim of isolating retroelements (*copia*, *gypsy*, LINE) where the PCRs are run with degenerate primers to a conserved region in the RT sequence. It is possible that degenerate primers would pick up a larger variety of PRV-L sequences from different plants. The basis for degenerate primers in the RT region would be already isolated sequences, SotuIII-1 from this study, the TPV family of sequences, TVCV, PVCV and BSV, although BSV might be too different from the others. Glyn Harper (pers. com.) has tried to find regions of homology from a selection of pararetroviruses long enough to make primers. The sequences are rather heterogeneous but there are conserved areas in the RT/RNase H region. Degenerate primers could best be made within closely related sequences but it is important to maintain a degree of specificity so that other families of retroelements are not amplified. In addition to the RT region the repetitive region of TPV amplified from many plants in this study could have possibilities for primer design. It would also be important to have primers from several regions of the PRV-L sequences to better be able to estimate if whole sequences are present.

Having successful and reliable PCRs one should sample systematically from a wide variety of species including algae and fungi. An interesting point would be to find species not generating a positive product although it merely tells what the primer pairs can amplify and not necessary that the species is deficient of PRV-L sequences. Amplifications could be extended to insects and mammals, although probably plant PRV-L sequences are not present in animals, because of host range of the viruses and maybe because of the alternative proliferation of retroviruses.

6.1.2 Southern hybridization

The technique of transferring DNA, digested and size separated in a gel by electrophoresis, to a membrane was first described by Southern (1975) and built on the principle of capillary transfer which moves DNA fragments from the gel to the membrane in unchanged order. The technique is widely used to identify specific sequences or sequences within digests of total genomic DNA. Southern hybridization can tell something about the presence of a sequence and to some extent the copy number, but further interpretation about whether a sequence is located in a few discrete clusters in the genome or present throughout the genome is limited. The drawback compared to PCR is the large amount of DNA (5-10 μg per lane) necessary from each accession but the advantages over *in situ* hybridization is that a large number of species can be screened comparatively easily.

Southern hybridization with PRV-like probes found the sequences present in genomic DNA of potato, other Solanaceous plants and a wider variety of plants from liverwort to Angiosperms. Although suspicion could be raised that some of the Southern hybridization result at low stringency might be the result of non-specific hybridization, therefore more tests would be necessary with other probes and varying stringencies. It would be useful to test other regions of a PRV-L sequence as probes to genomic DNA as here two regions were hybridized to potato and one region to other plants. An optimising factor for uniform results would be to estimate the DNA concentration before loading the gels.

6.1.3 *In situ* hybridization

Work such as that presented here also shows the important role of *in situ* hybridization in understanding genome organization in plant species. Unlike PCR methods, hybridization is much more free of artefacts, and can clearly show aspects of the chromosome-specificity and physical position which cannot easily by found by Southern hybridization, and demonstrates that sequences are nuclear and not present in organellar genomes or as episomal viral particles or other pathogen containing DNA. *In situ* hybridization including chromosome spread preparations are time consuming and rely on a lot of factors to work. Important things are good plant growth, root activity, a high metaphase index and well-labelled probes.

Here several different PRV-L probes were found to hybridize to potato metaphases, with several signals on most of the chromosomes. Even though two other species were tested here for the presence of PRV-L sequences, more thorough tests should be conducted on these and other plant species with various probes and different hybridization stringencies. It is also important to more consistently check which products and methods are working and which are not. Especially to optimise metaphase index and get better measure for probe quality.

6.1.4 Bioinformatics

The recent flood of data from genome sequencing and functional genomics has given rise to a new field, bioinformatics, which combines elements of biology and computer science. The recent years have seen large extensions of DNA, EST and protein sequence databases mainly because of large scale sequencing efforts in plant and animals (Lin *et al.*, 1999; Yu *et al.*, 2002; Goff *et al.*, 2002; International Human Genome Sequencing Consortium, 2001; Gregory *et al.*, 2002). This information gives rise to many, largely computer based, investigations of function and diversity of fragments in the genomes of various organisms. An extensive part of this study has also relied on database searching and generation of information from the internet. Obtaining, comparing and using other peoples sequences was the basis of the retroelement alignment in chapter I. Besides, the basic sequence analysis, alignment, phylogenetic tree generation, dotplotting etc. is conveniently performed by the computer.

The general aim of this study was to investigate the presence, nature and organization of PRV-L sequences integrated into the genome of potato and other plants. The detailed results of the above findings were described and discussed in the chapters, III, IV and V. The results will now be discussed in the broader context of other work.

6.2 Organization of PRV-L sequences in the host genome

6.2.1 Copy number

Jakowitsch *et al.* (1999) estimated the TPV-like sequences in tobacco by slot blot analysis to have a copy number of about 1000 per diploid genome. They detected the TPV by Southern but not by *in situ* hybridization and found most sequences to be more or less truncated. The sequence and hybridization data presented here show that it is difficult to measure copy number of PRV-like sequences because of the high inter-sequence variation and presence of partial sequence fragments. As discussed in chapter III, the primers will select only a subsection of all the genomic sequences present. Still this study makes a major advance towards better understanding the organization and copy number of PRV-L sequences in potato, though it is still not clear how large they are, and if they are complete or truncated.

In the experiments presented here potato genomic DNA gave strong PCR amplification with primers derived from the end and a middle part of TPV but whether the two parts are able to form a complete virus-like sequence are not clear. There may be both longer and smaller fragments present as it was possible to PCR amplify longer sequences of several kb. Southern hybridization to genomic DNA with PRV-like probes gave strong signals but metaphase *in situ* hybridization gave weak and unstable signal, which may indicate dispersal of multiple single to tandem copy sequences. Schwarzacher and Heslop-Harrison (2000) estimate that *in situ* hybridization probes can detect genome targets down to as small as around 1 kb but with increasing uncertainty as size reduces. The length of the probes used here was about 1 kb (fragmented) so the faint signals seen indicate there cannot be many copies of that sequence at any one location in the potato genome. Potato was the only one of a few Solanaceous species analysed which gave entirely satisfactory *in situ* hybridization results, but was also used for more experiments. Copy number and sequence length estimations could be more efficiently done using BAC library screening followed by sequencing. BAC libraries are publicly available for *Solanum pinnatisectum*, a wild Mexican diploid potato species (Chen and Zhang, 2001) and additionally Song *et al.* (2000) constructed a BAC library from diploid *Solanum bulbocastanum*.

The copy number of extracted PRV-L sequences could also be estimated using slotblot (Pearce *et al*, 1996; Bejarano *et al.*, 1996; Jakowitsch *et al.*, 1999) or real time PCR. Microarray technique could be used to scan a large number of plants, varying many factors at the same time. A long-term approach would be to sequence the whole genome.

There are not many PRV-L sequences in non-virus genome databases. After the whole *Arabidopsis* genome was sequenced there was no detectable PRV-L sequences. *Arabidopsis* has a very small genome and might lack repetitive DNA found in plants with larger genomes or such sequences may be short and well distributed in the genome making them difficult to identify. The focus of sequencing projects is often on genes and not on repetitive sequences which can also be more difficult to sequence.

6.2.2 Genome specificity

In some cases it is found that the integrated PRV-L sequences are restricted to part of the genome of an allopolyploid or to one parent of a hybrid.

Geering *et al.* (2001) used a Southern blot hybridization assay to examine many *Musa* cultivars (having many variations of A and B genomes) for the presence of an integrated sequence of BSV (BSV-OL) and found that this sequence was linked to the B genome. They also found a slightly different integrated virus in an AAA cultivar ('Williams'), which hybridized positively to DNA of other A genome containing cultivars. The results indicated that different *Badnavirus* integrants are restricted in their distribution to certain species or even subspecies of *Musa*.

Nicotiana tabacum is an allotetraploid hybrid between *N. sylvestris* and *N. tomentosiformis* or *N. otophora* donating the S and the T genome respectively to tobacco. In this study a specific *copia* retroelement hybridized more strongly to half of the chromosomes indicating that either S or T genome progenitors donated a higher copy number of this particular *copia* sequence. A gradient of a *copia* retroelement has also been seen in hexaploid oat, *Avena sativa* (Katsiotis *et al.*, 1996) with ACD genome. A *copia* probe derived from a C genome diploid hybridized more strongly to one third of the oat chromosomes.

It seems that allopolyploid species are very slow to recombine the DNA between different genomes probably because successful pairing of chromosomes has to be bivalents, keeping the crossing over within genomes. Perhaps certain retroelements persist in their original genome where host/element restrictions and expansions are in place. If the retroelement in question is not active anymore or only slightly active then only the pattern derived from the diploid genomes will be seen. Though one would think a foreign genome would be a possibility for quick expansion of an element before host restrictions were in place. No real bias was detected of PRV-L probes to potato chromosomes in this study and Hanson *et al.* (1999) found a *copia* retrotransposon dispersed on all chromosomes of allotetraploid cotton even though it was only originally present in one of the parents. So apparently there are different mechanisms of spread at work in different situations probably dependent on element type and host genome constitution.

6.3 Age

It is useful to know the age of integrated retroelements as it gives knowledge about how the host genome evolved. Is this during gradual expansion or short periods of high activity and is there a simultaneous removal of elements by the host genome? A burst of activity of retrotransposons was detected in the maize genome by SanMiguel *et al.* (1998).

The PRV-L elements found in the present study must date back to at least the division from algae, unless horizontal transfer has occurred, as it was found in *Marchantia* being later lost or very diminished in some lineages (e.g. *Arabidopsis*). The LTRs can be used to determine age in LTR containing elements, but it is difficult to know how to determine the age of inserted PRV-L sequences where the insertion mechanism is not clear. It may be possible to compare them with the closest episomal relative, in this case TVCV or CsVMV. But these viruses will themselves have evolved from the time of insertion. It is possible that the integration of BSV into the *Musa* genome is relatively recent as it is found only in some *Musa* germplasm (e.g. the B genome) although some homologous fragments have also been found in the A genome (Geering, 2001). Integration of pararetrovirus sequences in virus infected plants may be an ongoing process. Evidence has been found of CaMV integration in infected cells during natural infection, but there was no fixation detected (Al-kaff, 1994).

6.4 Integration

How could a pararetrovirus without integrase function become integrated in plant genomes in the first place? Alternatively, why are episomal DNA viruses, particularly those having nuclear location not frequently integrated into plant genomes? It is also known that genes, or large chromosome segments (e.g. in *Arabidopsis*, Lin *et al.*, 1999), are transferred from organellar to nuclear genomes, so mechanisms for integration of sequences must be present in plants. Alien DNA can also be integrated relatively easily by transformation techniques. Illegitimate integration of foreign DNA might happen more frequently than we imagine and could be transferred to offspring of readily vegetatively propagating plants such as potato and tomato. It may not occur readily in sexually propagating plants unless the integration happened in meristematic tissue. Integration of PRV-L sequences could be mediated by integrases from active retroelements, and Bennetzen (2000) has suggested that the action of the RT complex from retroelements can potentially turn any RNA (e.g. mRNA, virus in RNA state) into a DNA that can then be integrated into the genome. Features such as failure in gap repair, extended lifetime of open circular DNA, recombination between free ends of "virus" and plant DNA have been suggested by Jakowitsch *et al.* (1999) to contribute to integration. One pararetrovirus

PVCV might have genes for an integrase function (Richert-Pöggeler and Shepherd, 1997). The conserved motifs of integrase (HHCC and DD35E) were found in the middle of ORF1 of PVCV. PVCV may be a special pararetrovirus, and as the suggested position of the integrase motifs are different from that in other integrase containing elements (*copia*, *gypsy* and retroviruses), compared to the position of other conserved features, the virus could have acquired the integrase from another element by recombination. Viruses are also known to be very recombinogenic. Research found that when a virus (CaMV) deficient in a gene function was introduced into a plant containing a transgene for this function, the virus was able to get and utilize this fragment and complete the infection cycle (Király *et al.*, 1998). The same type of phenomenon has also been observed with a geminivirus (Frichmuth and Stanley, 1998). Perhaps some of the ORFs with unknown function in pararetroviruses contain coding for some form of integration.

This shows that there are several mechanisms by which fragments without their own integrase system can be integrated in nuclear DNA either by acquiring this function or by being worked upon by other elements.

6.5 Mobility within the genome

PRV-L sequences have been found dispersed in the genome of plants (this study; Lockhart and Schwarzacher, unpublished data (TVCV); Jakowitsch *et al.*, 1999, TPV). As pararetroviruses have no recognized integrase function it is not clear how they amplify in the genome. Unlike retroelements an integrated pararetrovirus sequence may not be able to amplify throughout the genome by itself. If the element or part of it is transcribed and reverse transcribed the trans-action from other elements could mediate integration as mentioned above. Other possibilities are that tandem duplications and rearrangements caused by unequal crossing over could expand and alter the sequence. Besides, other dispersed repeats with no known genomic spread mechanism are present in the genome of plants (e.g. Horáková and Fajkus, 2000).

6.6 Activity and expression

What is the degree of activity of integrated pararetroviruses? Are episomal forms generated under conditions such as stress or do they transcribe under normal conditions? Only very faint RNA transcripts of TPV and BSV were detected in tobacco and banana leaves respectively (Jakowitsch *et al.*, 1999; Harper and Hull, 1998) which could be due to host restrictions such as methylation. Considering that PRV-L sequences seem to be widespread in plants there is a limited number of ESTs in databases with high similarity to the endogenous viruses. This could possible be caused by lack of study and it might be that PRV-L sequences are mostly methylated and therefore transcription is absent or limited. Besides, not all virus-like transcripts have the required poly-A ending to be selected in extraction of mRNA for ESTs. In this case it might be difficult to get a good estimate for number of expressed PRV-L sequences in plant genomes. Perhaps immunoblotting could be used to detect translation products. For that isolated proteins have to be size separated in a gel and blotted to a membrane which can be probed with an antibody from a PRV-L translation product. Transcription of PRV-L sequences could also be analysed by isolation of RNA and Northern blot analysis (Pearce *et al.*, 1997). Sampling needs to be from different parts of the plant (including embryo and gametophyte) as there might be tissue specific expression and they might be expressed under different conditions such as varying levels of stress and ages.

Many transposable elements are found to be more or less degenerate and contain stop codons preventing transcription of functional proteins. Even if elements are degenerate, it is possible that some still functional part could be activated and processed by promoters and proteins from other elements or by a read through mechanism.

Mhiri *et al.* (1997, 1999) showed that various biotic and abiotic stresses (wounding, freezing or poisoning) activated the expression of the tobacco retrotransposon *Tnt1* in tobacco and also in tomato and *Arabidopsis*. There seemed to be a correlation between the onset of *Tnt1* expression and expression of plant defence genes. Some homology was found between the retroelement LTR and motifs involved in defence gene activation. It was not clear if the activation of *Tnt1* had any benefits for the plant in reducing the infection or giving relief of stress. Maybe pathogens or pathogen responses trigger retroelement transcription with the potential chance to get transferred to the pathogen or the insect vector and thereby the possibility of being spread to a new host. Additionally, also under natural conditions, environmental stress was found to have promoted the expansion of a retroelement (*BARE*-1) in *Hordeum* (Kalendar *et al.*, 2000).

Many repetitive sequences including PRV-L sequences are probably methylated by the plant to prevent transcription of harmful products and amplification in the genome. Perhaps the methylation is removed during pathogen infection as a stress reaction or because some sequences are beneficial in that situation. While direct evidence is not known it is worthwhile to speculate about the mutualism between PRV-L sequences and their host genome in order to develop models of behaviour to test.

6.7 Conservation

Surprisingly it has been found that retroelements and other repetitive sequences are more conserved between species or even distant families than would be expected of the genes of the same individuals (this study; Flavell *et al.*, 1992b; Wright and Voytas, 2002; Hughes, 2000; Teo *et al.*, 2002). Groups of *copia* retroelements isolated from various banana species and cultivars had up to 82% identity at the nucleotide level with a *copia* fragment from potato (Teo *et al.*, 2002). Similar high identities was found in the present study where PRV-L fragments from one banana cultivar ('Obino l'Ewai', MuObI-1 and MuObI-5) had a similarity ranging from 79-99% to corresponding potato fragments, though the ends of fragments are selected to be similar by using conserved PCR primers.

Benit *et al.* (1999) discuss why the ERV-L elements with widespread dispersion among mammals have a high level o f sequence conservation, usually a characteristic of a classical gene. One conclusion was that there is a continuous replacement of excised or mutated transposable elements by newly transposed copies from functional ERV-L elements, so an equilibrium is maintained between the insertion of new copies and the removal of old ones. This could account for the high level of sequence conservation. Alternatively the integrating elements have some kind of ability to maintain unchanged, or have some sort of function in the cell which promotes strong conservation.

6.8 Function

How could a PRV-L sequence survive in the plant genome? Supposedly by avoiding host decline through transcription of harmful viral products. But if the right control is maintained virus transcripts might be used to counteract new viruses through post transcriptional gene silencing (PTGS) (e.g. Matzke *et al.*, 2000). The trigger of PTGS is a short dsRNA sequence, which could derive from sense and anti-sense fragments, or sequences with inverted repeats as seen in PRV-L sequences (figure 3.11 and 3.12). Mette *et al.* (2002) isolated an enhancer region from TPV and showed that when reintroduced as a transgene it was silenced by the homologous endogenous form. It was from the same region of TPV that primers of the present study, which amplify fragments from many plant families, were derived.

So there are reasons to suppose that the presence of a conserved virus sequence in plants may have a function in silencing incoming virus although the PTGS function is not particularly clear. Some evidence points out that the plant endogenous sequence needs to be 100% homologous to the sequences in the infecting virus to have a strong effect (Carmichael, 2002; Jacque *et al.*, 2002). Given that viruses are known to be rapidly changing it is not clear how this mechanism would work unless there are short (e.g. 21-25 bp) 100% conserved regions present. The study of Mette *et al.* (2000) was performed with a self-transgene which then ought to be 100% homologues.

Virus-like fragments could also have some regulatory function in the cell. Viruses have strong promoters (e.g. CaMV 35S promoter) and some also have transactivators (TPV, TVCV) which inserted before genes in the host genome could give new or at least altered patterns of protein production in a cell.

6.9 Significance for plant genome evolution

Polyploidization is an important factor in the evolution of plants and can give changes to genetic and genomic aspect of the hybrids for instance nucleolar dominance, gene silencing and activation of transposable elements. Transposable elements have on their part over evolutionary time increased the diversity of the genome through expansion, mutation and transactivation.

Many experiments and much speculation have been directed towards the organization and evolution of *gypsy* and especially *copia* retrotransposable elements (Flavell *et al.*, 1997; Heslop-Harrison *et al*, 1997; Kumar *et al.*, 1997; Friesen *et al.*, 2001), but integrated pararetroviruses are a new field. Complete PRV-L elements or elements able to replicate have been found in the Solanaceae and *Musa* species. These species are not actually closely related one being a Dicotyledon the other a

Monocotyledon. These might be chance findings as the present study indicates PRV-L sequences to be wide spread in plant genomes. The problem with integrated BSV sequences in *Musa* has become a significant problem for the breeding in the three years since its discovery. The same could happen in other crop or ornamental plants if climatic factors are changed, new crosses investigated or if tissue culture are introduced. Plants used in the present study were not apparently visually infected by any virus so maybe they contain harmless sequences unable to be activated, chronic unrecognized infection or the trigger for activation has not been present. More results are needed to confirm and extend the results of the present study to gain more knowledge of integrated PRV-L sequences behaviour and consequences for the host.

Future research in the topic will show the extent and the consequences of the presence of PRV-L sequences in genomes and give insight into how these elements spread. The data may show how the host organisms utilize PRV-L sequences for pathogen defence and stress responses, and if PRV related sequences may one day become a valuable tool in plant genetic enhancement.

References

Al-Kaff NS, 1994. Biological and molecular diversity of *Cauliflower mosaic virus*. PhD thesis, University of East Anglia.

Ashby MK, Warry A, Bejarano ER, Khashoggi A, Burrell M, Lichtenstein CP, 1997. Analysis of multiple copies of geminiviral DNA in the genome of four closely related *Nicotiana* species suggests a unique integration event. Plant Mol Biol 35: 313-321.

Auge-Gouillou C, Bigot Y, Pollet N, Hamelin M-H, Meunier-Rotival M, Periquet G, 1995. Human and other mammalian genomes contain transposons of the *mariner* family. FEBS Lett 368: 541-546.

Azpiroz-Leehan R, Feldmann KA, 1997. T-DNA insertion mutagenesis in *Arabidopsis* going back and forth. Trends Genet 13: 152-156.

Baltimore D, 1970. RNA-dependant DNA polymerase in virions of RNA tumour viruses. Nature 226: 1209-1211.

Bejarano ER, Khashoggi A, Witty M, Lichtenstein C, 1996. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. PNAS 93: 759-764.

Bell PJL, 2001. Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? J Mol Evol 53: 251-256.

Bénit L, Lallemand J-B, Casella J-F, Philippe H, Heidmann T, 1999. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. J Virol 73: 3301-3308.

Bennett MD, Leitch IJ, 1995. Nuclear DNA amounts in angiosperms. Ann Botany 76: 113-176.

Bennett MD, Leitch IJ, 1997. Nuclear DNA amounts in angiosperms - 583 new estimates. Ann Botany 80: 169-196.

Bennett MD, Leitch IJ, 2001. Nuclear DNA amounts in Pteridophytes. Ann Botany 87: 335-345.

Bennett MD, Smith JB, 1991. Nuclear DNA amounts in angiosperms. Philos Trans R Soc London B 334: 309-345.

Bennett MD, Bhandol P, Leitch IJ, 2000. Nuclear DNA amounts in angiosperms and their modern uses - 807 new estimates. Ann Botany 86: 859-909.

Bennetzen JL, 2000. Transposable element contribution to plant genome evolution. Plant Mol Biol 42: 251-269.

Bestor TH, 1999. Sex brings transposons and genomes into conflict. Genetica 107: 289-295.

Blackburn EH, 1992. Telomerases. Annu Rev Biochem 61: 113-129.

Blackburn EH, 2000. The end of the (DNA) line. Nature Struct Biol 7: 847-850.

Boeke JD, Corces VG, 1989. Transcription and reverse transcription of retrotransposons. Annu Rev Microbiol 43: 403-434.

Boeke JD, Eichinger D, Castrillon D, Fink GR, 1988. The *Saccharomyces cerevisiae* genome contains functional and nonfunctional copies of transposon Tyl. Mol Cell Biol 8: 1432-1442.

Brandes A, Heslop-Harrison JS, Kamm A, Kubis S, Doudrick RL, Schmidt T, 1997. Comparative analysis of the chromosomal and genomic organization of Tyl-*copia*-like retrotransposons in pteridophytes, gymnosperms and angiosperms. Plant Mol Biol 33:11-21.

Budiman MA, Mao L, Wood TC, Wing RA, 2000. A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. Genome Res 10: 129-136.

Capy P, Anxolabéhère D, Langin T, 1994. The strange phylogenies of transposable elements: are horizontal transfer the only explanation? Trends Genet 10: 7-12.

Capy P, Italis R, Langin T, Higuet D, Bazin C, 1996. Relationship between transposable elements based upon the integrase-transposase domains: Is there a common ancestor? J Mol Evol 42: 359-368.

Capy P, Langin T, Higuet D, Maurer P, Bazin C, 1997. Do the integrase of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100: 63-72.

Carmichael GG, 2002. Silencing viruses with RNA. Nature 418: 379-380.

Chavanne F, Zhang D-X, Liaud M-F, Cerff R, 1998. Structure and evolution of *Cyclops*: a novel giant retrotransposon of the *Ty3/Gypsy* family highly amplified in pea and other legume species. Plant Mol Biol 37: 363-375.

Chen Q, Zhang H, 2001. Construction of two BAC libraries from a wild Mexican diploid potato: *Solanum pinnatisectum*. Abstract, poster session, Plant and Animal Genome IX Conference, San Diego, Jan 13-17.

Clare J, Farabaugh P, 1985. Nucleotide sequence of a yeast Ty element: Evidence for an unusual mechanism of gene expression. PNAS 82: 2829-2833.

Covey SN, 1986. Amino acids sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of *Cauliflower mosaic virus*. Nucleic Acids Res 14: 623-633.

Dahal G, Hughes J d'A, Thottappilly, 1998. Effect of temperature on symptom expression and reliability of Banana streak badnavirus detection in naturally infected plantain and banana (*Musa* sp.). Plant Disease 82: 16-21.

Dahal G, Ortiz R, Tenkouano A, Hughes J d'A, Thottappilly G, Vuylsteke D, Lockhart BEL, 2000. Relationship between natural occurrence of banana streak badnavirus and symptom expression, relative concentration of viral antigen, and yield characteristics of some micropropagated *Musa* spp. Plant Pathology 49: 68-79.

Dai L, Zimmerly S, 2002. Compilation and analysis of group II intron insertion in bacterial genomes: evidence for retroelement behaviour. Nucleic Acids Res 30: 1091-1102.

Dallot S, Acuña P, Rivera C, Ramírez P, Côte F, Lockhart BEL, Caruana ML, 2001. Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of *Banana streak virus* integrated into the genome of the FHIA 21 hybrid (*Musa* AAAB). Arch Virol 146: 2179-2190.

Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A, 1990. Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species. Genetics 124: 339-355.

Dong F, Song J, Naess SK, Helgeson JP, Gebhardt C, Jiang J, 2000. Development and applications of a set of chromosome-specific cytogenetic DNA markers in potato. Teor Appl Genet 101: 1001-1007.

Eickbush TH, 1997. Telomerase and retrotransposons: Which came first? Science 277: 911-912.

Eickbush TH, 1999. Mobile introns: Retrohoming by complete reverse splicing. Curr Biol 9: dispatch R11-R14.

Emori Y, Shiba T, Kanaya S, Inouye S, Yuki S, Saigo K, 1985. The nucleotide sequence of *copia* and *copia*-related RNA in *Drosophila* virus-like particles. Nature 315: 773-776.

Fayet O, Ramond P, Polard P, Prère MF, Chandler M, 1990. Functional similarities between retroviruses and the IS*3* family of bacterial insertion sequences? Mol Microbiol 4: 1771-1777.

Feng Q, Moran JV, Kazazian HH, Boeke JD, 1996. Human L1 retrotransposons encodes a conserved endonuclease required for retrotransposition. Cell 87: 905-916.

Feschotte C, Wessler SR, 2002. *Mariner*-like transposases are widespread and diverse in flowering plants. PNAS 99: 280-285.

Feuerbach F, Drouaud J, Lucas H, 1997. Retrovirus-like end processing of the tobacco Tnt1 retrotransposon linear intermediates of replication. J Virol 71: 4005-4015.

Finnegan DJ, 1985. Transposable elements in eukaryotes. Int Rev Cytol 93: 281-326.

Flavell AJ, Smith DB, 1992. A *Tyl-copia* group retrotransposon sequence in a vertebrate. Mol Gen Genet 233: 322-326.

Flavell AJ, Dunbar E, Anderson R, Pearce SR, Hartley R, Kumar A, 1992a. *Tyl-copia* group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Res 20: 3639-3644.

Flavell AJ, Smith DB, Kumar A, 1992b. Extreme heterogeneity of *Tyl-copia* group retrotransposons in plants. Mol Gen Genet 231: 233-242.

Flavell AJ, Pearce SR, Heslop-Harrison P, Kumar A, 1997. The evolution of Ty1-*copia* group retrotransposons in eukaryote genomes. Genetica 100: 185-195.

Franck A, Guilley H, Jonard G, Richards K, Hirth L, 1980. Nucleotide sequence of cauliflower mosaic virus DNA. Cell 21: 285-294.

Friesen N, Brandes A, Heslop-Harrison JS, 2001. Diversity, origin and distribution of retrotransposons (*gypsy* and *copia*) in conifers. Mol Biol Evol 18: 1176-1188.

Frischmuth T, Stanley J, 1998. Recombination between viral DNA and the coat protein gene of *African cassava mosaic virus*. J Gen Virol 79: 1265-1271.

Gawel NJ, Jarret RL, 1991. A modified CTAB DNA extraction procedure for *Musa* and *Ipomea*. Plant Mol Biol Rep 9: 262-266.

Geering ADW, Olszewski NE, Dahal G, Thomas JE, Lockhart BEL, 2001. Analysis of the distribution and structure of integrated *Banana streak virus* in a range of *Musa* cultivars. Mol Plant Pathol 2: 207-213.

Gerlach WL, Bredbrook JR, 1979. Cloning and characterization of ribosomal RNA genes from wheat and barley. Nucleic Acids Res 7: 1869-1885.

Gerlach WL, Dyer TA, 1980. Sequence organization of the repeating units in the nucleus of the wheat which contain 5S rRNA genes. Nucleic Acids Res 8: 4851-4855.

Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, *et al*, 2002. A draft sequence of the Rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296: 92-100.

Gough KC, Cockburn W, Whitelam GC, 1999. Selection of phage-display peptides that bind to cucumber mosaic virus coat protein. J Virol Methods 79: 169-180.

Grandbastien M-A, 1998. Activation of plant retrotransposons under stress conditions. Trends Plant Sci 3: 181-187.

Gregor W, Mette MF, van der Winden J, Matzke AJ, Matzke MA, 2002. Endogenous pararetrovirus-like sequences in *Nicotiana otophora* and *Nicotiana tomentosiformis*. Unpublished data.

Gregory SG, Sekhon M, Schein J, Zhao SY, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, *et al*, 2002. A physical map of the mouse genome. Nature 418: 743-750.

Hall BG, 2001. Phylogenetic trees made easy. A how-to manual for molecular biologists. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, USA.

Hanson RE, Nurul Islam-Faridi M, Crane CF, Zwick MS, Czeschin DG, Wendel JF, McKnight TD, Price HJ, Stelly DM, 2000. Ty1-*copia*-retrotransposon behavior in a polyploid cotton. Chromosome Res 8: 73-76.

Harper G, Hull R, 1998. Cloning and sequence analysis of banana streak virus DNA. Virus Genes 17: 271-278.

Harper G, Osuji JO, Heslop-Harrison JS, Hull R, 1999. Integration of banana streak badnavirus into the *Musa* genome: Molecular and cytogenetic evidence. Virology 255: 207-213.

Harper G, Hull R, Lockhart B, Olszewski N, 2002. Viral sequences integrated into plant genomes. Annu Rev Phytopathol 40: 119-136.

Hawkes JG, 1990. The potato Evolution, biodiversity and genetic resources. Belhaven Press, London.

Herniou E, Martin J, Miller K, Cook J, Wilkinson M, Tristem M, 1998. retroviral diversity and distribution in vertebrates. J Virology 72: 5955-5966.

Heslop-Harrison JS, 2000. RNA, genes, genomes and chromosomes: Repetitive DNA sequences in plants. Chromosomes Today 13: 45-56.

Heslop-Harrison JS, Brandes A, Taketa S, Schmidt T, Vershinin AV, Alkhimova EG, Kamm A, Doudrick RL, Schwarzacher T, Katsiotis A, Kubis S, Kumar A, Pearce SR, Flavell AJ, Harrison GE, 1997. The chromosomal distribution of Tyl-*copia* group retrotransposable elements in higher plants and their implication for genome evolution. Genetica 100: 197-204.

Hirochika H, Hirochika R, 1993. Tyl-copia group retrotransposons as ubiquitous components of plant genomes. Jap J Genet 68: 35-46.

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M, 1996. Retrotransposons of rice involved in mutations induced by tissue culture. PNAS 93: 7783-7788.

Hohn T, Fütterer J, 1997. The proteins and functions of plant pararetroviruses: knowns and unknowns. Crit Rev Plant Sci 16: 133-161.

Horáková M, Fajkus J, 2000. TAS49 – a dispersed repetitive sequence isolated from subtelomeric regions of *Nicotiana tomentosiformis* chromosomes. Genome 43: 273-284.

Hughes DC, 2000. MIRs as agents of mammalian gene evolution. Trends Genet 16: 60-62.

Hull R, 1999. Classification of reverse transcribing elements: a discussion document. Arch Virol 144: 209-214.

Hull R, 2002. Plant virology. Academic Press, London, UK.

Hull R, Covey SN, 1996. Retroelements: Propagation and adaptation. Virus Genes 11: 105-118.

Index to plant chromosome numbers. 1975 –. Missouri Botanical Garden. USA

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Jacobsen N, Jensen J, 1997. Systematisk botanik (Systematic botany). DSR Forlag, Copenhagen.

Jacque J-M, Triques K, Stevenson M, 2002. Modulation of HIV-1 replication by RNA interference. Nature 418: 435-438.

Jakowitsch J, Mette MF, van der Winden J, Matzke MA, Matzke AJM, 1999. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. PNAS 96: 13241-13246.

Jordan IK, Matyunina LV, McDonald JF, 1999. Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. PNAS 96: 12621-12625.

Judd WS, Campell CS, Kellogg EA, Stevens PF, 1999. Plant systematics a phylogenetic approach. Sinauer Associates, Inc., Publishers Sunderland, Massachusetts USA.

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH, 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. PNAS 97: 6603-6607.

Katsiotis A, Schmidt T, Heslop-Harrison JS, 1996. Chromosomal and genomic organization of Ty1-*copia*-like retrotransposons sequences in the genus *Avena*. Genome 39: 410-417.

Kenton A, Khashoggi A, Parokonny A, Bennett MD, Lichtenstein C, 1995. Chromosomal location of endogenous geminivirus related DNA sequences in *Nicotiana tabacum* L. Chromosome Res 3: 346-350.

Khan E, Mack JPG, Katz RA, Kulkosky J, Skalka AM, 1991. Retroviral integrase domains: DNA binding and the recognition of LTR sequences. Nucleic Acids Res 19: 851-860.

Király L, Bourque JE, Schoelz JE, 1998. Temporal and spatial appearance of recombinant viruses formed between *Cauliflower mosaic virus* (CaMV) and CaMV sequences present in transgenic *Nicotiana bigelovii.* Mol Plant-Microbe Interact 11: 309-316.

de Kochko A, Verdaguer B, Taylor N, Carcamo R, Beachy RN, Fauquet C, 1998. Cassava vein mosaic virus (CsVMV), type species for a new genus of plant double stranded DNA viruses? Arch Virol 143, 945-962.

Kubis S, Schmidt T, Heslop-Harrison JS, 1998. Repetitive DNA elements as a major component of plant genomes. Ann Botany 82 (Supp A): 45-55.

Kubis SE, Castilho AMMF, Vershinin AV, Heslop-Harrison J-S, 2002. Retroelements, transposons and methylation status in the genome of oil palm (*Elaeis guineensis*) and the relationship to somoclonal variation. Plant Mol Biol, in press.

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

Kumar A, Pearce SR, McLean K, Harrison G, Heslop-Harrison JS, Waugh R, Flavell AJ, 1997. The Ty1-*copia* group of retrotransposons in plants: genomic organisation, evolution, and use as molecular markers. Genetica 100: 205-217.

Kumekawa N, Ohtsubo E, Ohtsubo H, 1999. Identification and phylogenetic analysis of *gypsy*-type retrotransposons in the plant kingdom. Genes Genetic Syst 74: 299-307.

Kunze R, Saedler H, Lönning W-E, 1997. Plant transposable elements. Adv Bot Res 27: 331-470.

Lawrence E, 1988. Henderson's dictionary of biological terms, 10th edition. Longman Scientific and Technical.

Le Q H, Wright S, Yu Z, Bureau T, 2000. Transposon diversity in *Arabidopsis thaliana*. PNAS 97: 7376-781.

Lecellier C-H, Saïb A, 2000. Foamy viruses: Between retroviruses and pararetroviruses. Virology 271: 1-8.

Leitch IJ, Chase MW, Bennett MD, 1998. Phylogenetic analysis of DNA C-value. Ann Botany 82 (Supp. A): 85-94.

Leitch IJ, Hanson L, Winfield M, Parker J, Bennett MD, 2001. Nuclear DNA C-values complete familial representation in Gymnosperms. Ann Botany 88: 843-849.

Lerat E, Brunet F, Bazin C, Capy P, 1999. Is the evolution of transposable elements modular? Genetica 107: 15-25.

Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, *et al*, 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402: 761-768.

Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR, 1997. Reverse transcriptase motifs in the catalytic subunit of telomerase. Science 276: 561-567.

Lockhart BE, Menke J, Dahal G, Olszewski NE, 2000. Characterization and genomic analysis of tobacco vein-clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. J Gen Virol 81: 1579-1585.

Löwer R, 1999. The pathogenic potential of endogenous retroviruses: facts and fantasies. Trends Microbiol 7: 350-356.

Löwer R, Löwer J, Kurth R, 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. PNAS 93: 5177-5184.

Malik HS, Eickbush TH, 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res 11: 1187-1197.

Malik HS, Burke WD, Eickbush TH, 1999. The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 16: 793-805.

Malik HS, Henikoff S, Eickbush TH, 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10: 1307-1318.

Maniatis T, Fritsch EF, Sambrook J, 1982. Molecular cloning. A laboratory manual. Cold Spring Harbor, New York.

Manninen I, Schulman AH, 1993. *BARE-1*, a *copia*-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol Biol 22:829-846.

Mao L, Wood TC, Yu Yeisoo, Budiman MA, Tomkins J, Woo S-S, Sasinowski M, Presting G, Frisch D, Goff S, Dean RA, Wing RA, 2000. Rice transposable elements: a survey of 73.000 sequence-tagged-connectors. Genome Res 10: 982-990.

Marín I, Lloréns C, 2000. *Ty3/Gypsy* retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. Mol Biol Evol 17: 1040-1049.

Mathews DH, Sabina J, Zuker M, Turner DH, 1999. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. J Mol Biol 288: 911-940.

Matzke M, Matzke AJM, Kooter JM, 2001. RNA: Guiding gene silencing. Science 293: 1080-1083.

Matzke MA, Mette MF, Matzke AJM, 2000. Transgene silencing by the host genome defence: implications for the evolution of epigenetic control mechanisms in plants and vertebrates. Plant Mol Biol 43: 401-415.

McClintock B, 1984. The significance of responses of the genome to challange. Science 226: 792-801.

McClure MA, 1991. Evolution of retrotransposons by acquisition or deletion of retrovirus-like genes. Mol Biol Evol 8: 835-856.

Mette MF, Kanno T, Aufsatz W, Jakowitsch J, van der Winden J, Matzke MA, Matzke AJM, 2002. Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. EMBO J 21: 461-469.

Mhiri C, Morel J-B, Vernhettes S, Casacuberta JM, Lucas H, Grandbastien M-A, 1997. The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. Plant Mol Biol 33: 257-266.

Mhiri C, de Wit PJGM, Grandbastien M-A, 1999. Activation of the promoter of the Tnt1 retrotransposon in tomato after inoculation with the fungal pathogen *Clodosporium fulvum*. Mol Plant-Microbe Interact 12: 592-603.

Miller AD, 1997. Development and applications of retroviral vectors. In Retroviruses, Coffin JM, Hughes SH, Varmus HE, (ed.) p. 757; Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, NY, USA.

Miller K, Lynch C, Martin J, Herniou E, Tristan M, 1999. Identification of multiple *Gypsy* LTR-retrotransposon lineages in vertebrate genomes. J Mol Evol 49: 358-366.

Murray BG, 1998. Nuclear DNA amounts in Gymnosperms. Ann Botany 82 (Supp. A): 3-15.

Nasu M, Tani K, Hattori C, Honda M, Shimaoka T, Yamaguchi N, Katoh K, 1997. E fficient transformation of *Marchantia polymorpha* that is haploid and has very small genome DNA. J Ferment Bioengineer 84: 519-523.

Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski NE, Lockhart B, 1999. Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255: 214-220.

Olszewski N, Hagen G, Guilfoyle TJ, 1982. A transcriptionally active, covalently closed minichromosome of cauliflower mosaic virus DNA isolated from infected turnip leaves. Cell 29: 395-402.

Ono M, 1986. Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes. J Virol 58: 937-944.

Page RDM, 1996. TreeView: An application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357-358.

Panstruga R, Büschges R, Piffanelli P, Schulze-Lefert P, 1998. A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res 26: 1056-1062.

Pardue ML, Danilevskaya ON, Traverse KL, Lowenhaupt K, 1997. Evolutionary links between telomeres and transposable elements. Genetica 100: 73-84.

Pauk J, Stefanov I, Fekete S, Bogre L, Karsai I, Feher A, Dudidts D, 1995. A study of different CaMV 35S and MAS promoter activities and risk assessment of field use in transgenic rapeseed plants. Euphytica 85: 411-416.

Pearce SR, Harrison G, Li D, Heslop-Harrison JS, Kumar A, Flavell AJ, 1996. The *Tyl-copia* group retrotransposons in *Vicia* species: copy number, sequence heterogeneity and chromosome localisation. Mol Gen Genet 250: 305-315.

Pearce SR, Harrison G, Heslop-Harrison P, Flavell AJ, Kumar A, 1997. Characterization and genomic organization of Tyl-*copia* group retrotransposons in rye (*Secale cereale*). Genome 40: 617-625.

Petropoulos CJ, 1997. Appendix 2: Retroviral taxonomy, protein structure, sequences, and genetic maps. In Retroviruses, Coffin JM, Hughes SH, Varmus HE, (ed.), p. 757; Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, NY, USA.

Pfeiffer P, Hohn T, 1983. Involvment of reverse transcriptase in the replication of cauliflower mosaic virus: a detailed model and test of some aspects. Cell 33: 781-789.

Pineau P, Marchio A, Terris B, Mattei M-G, Tu Z-X, Tiollais P, Dejean A, 1996. A t(3;8) chromosomal translocation associated with Hepatitis B virus integration involves the carboxypeptidase N locus. J Virol 70: 7280-7284.

Pringle CR, 1999. Virus taxonomy. Arch Virol 144: 421-429.

Pustell J, Kafatos FC, 1982. A high speed, high capacity matrix: zooming through SV40 and polyoma. Nucleic Acids Res 10: 4765-4782.

Qiu Y-L, Cho Y, Cox JC, Palmer JD, 1998. The gain of three mitochondrial introns identifies liverworts as the earliest land plant. Nature 394: 671-674.

Qiu Y-L, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW, 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402: 404-407.

Renne R, Friedl E, Schweizer M, Fleps U, Turek R, Neumann-Haefelin D, 1992. Genomic organization and expression of simian foamy virus type 3 (SFV-3). Virology 186: 597-608.

Richards KE, Guilley H, Jonard G, 1981. Further charaterization of the discontinuities in cauliflower mosaic virus DNA. FEBS lett 134: 67-70.

Richert-Pöggeler KR, Shepard RJ, Caspar R, 1996. Petunia vein clearing virus, a pararetrovirus that exists as a retroelement in the chromosomes of its host. Abstracts of Xth International Congress of Virology, Jerusalem, Israel, W05-1.

Richert-Pöggeler KR, Shepherd RJ, 1997. Petunia vein-clearing virus: a plant pararetrovirus with the core sequence for an integrase function. Virology 236: 137-146.

Robertson HM, 1993. The *mariner* transposable element is widespread in insects. Nature 362: 241-245.

Saigo K, Kugimiya W, Matsuo Y, Inouye S, Yoshioka K, Yuki S, 1984. Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster.* Nature 312: 659-661.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N, 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of scile-cell anemia. Science 230: 1350-1354.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA, 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA-polymerase. Science 239: 487-491.

Saitou N, Nei M, 1987. The Neighbor-Joining method - a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL, 1998. The paleontology of intergene retrotransposons of maize. Nature Genet 20: 43-45.

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH, 1997. Many human L1 elements are capable of retrotransposition. Nature genet 16: 37-43.

Schenk PM, Remans T, Sági L, Elliott AR, Dietzgen RG, Swennen R, Ebert PR, Grof CPL, Manners JM, 2001. Promoters for pregenomic RNA of banana streak badnavirus are active for transgene expression in monocot and dicot plants. Plant Mol Biol 47: 399-412.

Schmidt T, 1999. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Mol Biol 40: 903-910.

Schwarzacher T, Heslop-Harrison P, 2000. Practical *in situ* hybridization. Bios, Oxford, UK.

Seeger C, 1999. Molecular Biology (of Hepatitis B). Encyclopedia of Virology, 2nd edition, ed. Granoff A, Webster RG. Academic Press, London.

Sharp PA, 2001. RNA interference – 2001. Genes Dev 15: 485-490.

Song J, Dong F, Jiang J, 2000. Construction of a bacterial artificial chromosome library for potato molecular cytogenetics research. Genome 43: 199-204.

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

Southern E, 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98: 503.

Staginnus C, Huettel B, Desel C, Schmidt T, Kahl G, 2001. A PCR-based assay to detect *En/Spm*-like transposon sequences in plants. Chromosome Res 9: 591-605.

Stuart-Rogers C, Flavell AJ, 2001. The evolution of Tyl-*copia* group retrotransposons in Gymnosperms. Mol Biol Evol 18: 155-163.

Suoniemi A, Tanskanen J, Schulman AH, 1998. *Gypsy*-like retrotransposons are widespread in the plant kingdom. Plant J 13: 699-705.

Tagieva NE, Gizatullin RZ, Zakharyev VM, Kisselev LL, 1995. A genome-integrated hepatitis B virus DNA in human neuroblastoma. Gene 152: 277-278.

Takahashi K, Akahane Y, Hino K, Otha Y, Mishiro, 1998. Hepatitis B virus genomic sequence in the circulation of hepatocellular carcinoma patients: comparative analysis of 40 full-length isolates. Arch Virol 143: 2313-2326.

Temin HM, Mizutani S, 1970. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature 226: 1211-1213.

Temsch EM, Greilhuber J, Krisai R, 1998. Genome size in *Spagnum* (peat moss). Botanica Acta 111: 325-330.

Teo CH, Tan SH, Othman YR, Schwarzacher T, 2002. The cloning of *Ty*l-*copia*-like retrotransposons from 10 varieties of banana (*Musa* sp.). J Biochem Mol Biol Biophys 6: 193-201.

Tillmann HL, Heiken H, Knapik-Botor A, Heringlake S, Ockenga J, Wilber JC, Goergen B, Detmer J, McMorrow M, Stoll M, Schmidt RE, Manns MP, 2001. Infection with GB virus C and reduced mortality among HIV-infected patients. New England J Med 345: 715-724.

Tzafrir I, Torbert KA, Lockhart BEL, Somers DA, Olszewski E, 1998. The sugarcane bacilliform badnavirus promoter is active in both monocots and dicots. Plant Mol Biol 38: 347-356.

Verdaguer B, de Kochko A, Fux CI, Beachy RN, Fauquet C, 1998. Functional organization of the cassava vein mosaic virus (CsVMV) promoter. Plant Mol Biol 37: 1055-1067.

Vershinin AV, Druka A, Alkhimova AG, Kleinhofs A, Heslop-Harrison JS, 2002. LINEs and *gypsy*-like retrotransposons in *Hordeum* species. Plant Mol Biol 49: 1-14.

Vicient CM, Suoniemi A, Anamthawat-Jónsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH, 1999. Retrotransposon *BARE-1* and its role in genome evolution in the genus *Hordeum*. The Plant Cell 11: 1769-1784.

Vicient CM, Kalendar R, Sculman AH, 2001. *Envelope*-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. Genome Res 11: 2041-2049.

Voglmayr H, 2000. Nuclear DNA amounts in mosses (*Musci*). Ann Botany 85: 531-546.

Wang P-C, Hui E K-W, Chiu J-H, Lo SJ, 2001. Analysis of integrated hepatitis B virus DNA and flanking cellular sequence by inverse polymerase chain reaction. Journal of Virol Methods 92: 83-90.

Waterhouse PM, Wang M-B, Lough T, 2001. Gene silencing as an adaptive defence against viruses. Nature 411: 834-842.

Waugh O'Neill RJ, O'Neill MJ, Marshall Graves JA, 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature 393: 68-72.

Wessler SR, 2001. Plant transposable elements. A hard act to follow. Plant Physiol 125: 149-151.

Wright DA, Voytas DF, 1998. Potential retroviruses in plants: *Tat1* is related to a group of *Arabidopsis thaliana Ty3/gypsy* retrotransposons that encode envelope-like proteins. Genetics 149: 703-715.

Wright DA, Voytas DF, 2002. *Athila4* of *Arabidopsis* and *Calypso* of Soybean define a linage of endogenous plant retroviruses. Genome Res 12: 122-131.

Xiang J, Wunschmann S, Diekema DJ, Klinzman D, Patrick KD, George SL, Stapleton JT, 2001. Effect of coinfection with GB virus C on survival among patients with HIV infection. New England J Med 345: 707-714.

Xiong Y, Eickbuch TH, 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353-3362.

Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ, et al, 2002. A draft of the Rice genome (*Oryza sativa* L. ssp. *indica*). Science 296: 79-92.

Zimmerly S, Hausner G, Wu X-C, 2001. Phylogenetic relationship among group II intron ORFs. Nucleic Acids Res 29: 1238-1250.

Zuker M, Mathews DH, Turner DH, 1999. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In RNA Biochemistry and Biotechnology, 11-43, Barciszewski J, Clark BFC, (ed.) NATO ASI Series, Kluwer Academic Publishers.