The EMBO Journal vol.14 no.11 pp.2670–2678, 1995

Sequence variability within the tobacco retrotransposon Tnt1 population

Josep M.Casacuberta[1,2], Samantha Vernhettes and Marie-Angèle Grandbastien

Laboratoire de Biologie Cellulaire, INRA, Centre de Versailles, 78026 Versailles, Cedex, France
[1]Present address: Departament de Genètica Molecular, Centre d'Investigació i Desenvolupament (CSIC), Jordi Girona 18, 08034 Barcelona, Spain

[2]Corresponding author

Communicated by T.Hohn

Retroviruses consist of populations of different but closely related genomes referred to as quasispecies. A high mutation rate coupled with extremely rapid replication cycles allows these sequences to be highly interconnected in a rapid equilibrium. It is not known if other retroelements can show a similar population structure. We show here that when the tobacco Tnt1 retrotransposon is expressed, its RNA is not a unique sequence but a population of different but closely related sequences. Nevertheless, this highly variable population is not in a rapid equilibrium and could not be considered as a quasispecies. We have thus named the structure presented by Tnt1 RNA quasispecies-like. We show that the expression of Tnt1 in different situations gives rise to different populations of Tnt1 RNA sequences, suggesting an adaptive capacity for this element. The analysis of the variability within the total genomic population of Tnt1 elements shows that mutations frequently occur in important regulatory elements and that defective elements are often produced. We discuss the implications that this population structure could have for Tnt1 regulation and evolution.
Key words: quasispecies/retrotransposon/tobacco/variability

Introduction

Replication of retroviruses is a complex process involving host- and virus-encoded enzymes. Two viral genomic RNAs transcribed by the host machinery are co-packaged in a viral particle where they are retrotranscribed by a virus-encoded reverse transcriptase to produce a single DNA copy of the element which, after infection of a new cell, will integrate into the genome (Varmus and Brown, 1989; Hu and Temin, 1990a). This unique mechanism of replication is very error-prone, due in particular to the lack of proofreading activity of the reverse transcriptase and the possibility of recombination during reverse transcription, and can generate a high degree of genomic heterogeneity (Hu and Temin, 1990b; Katz and Skalka, 1990; Temin, 1993). This genomic variability shared by most, if not all, retroviruses and RNA viruses is one of the reasons why some of these elements evolve at a high rate and, in some cases, escape the host's immune response, as has been proposed for HIV (Coffin, 1986; Meyerhans *et al.*, 1989; Wain-Hobson, 1993). Therefore, these viruses cannot be defined by a unique sequence, and it is necessary to describe them in terms of a population of closely related genomes, referred to as quasispecies (Domingo *et al.*, 1985; Steinhauer and Holland, 1987).

Retrotransposons are mobile genetic elements structurally and functionally closely related to retroviruses. Their central coding region is flanked by long directly repeated sequences (LTRs) and they transpose via an RNA intermediate which is reverse transcribed in virus-like particles (VLP) (Boeke and Corces, 1989). The homology between retroviruses and retrotransposons is so high that, based on comparisons of their reverse transcriptase sequences, it has been suggested that retroviruses could have evolved from a retrotransposon-like ancestor (Xiong and Eickbush, 1990). Nevertheless, retrotransposons lack an *env*-associated open reading frame (ORF) and they are supposedly non-infectious, their complete life cycle being intracellular. As retrotransposition closely resembles retroviral replication, it should also be expected to be very error-prone. It has been shown that mutations can be introduced during transposition of the yeast retrotransposon Ty1 (Boeke *et al.*, 1985), and that the yeast genome contains mutated Ty1 elements that have become inactive (Boeke *et al.*, 1988). Furthermore, the mutation rate during a single cycle of Ty1 transposition has been recently estimated to be 2.5×10^{-5} per nucleotide (Gabriel, Willems, Mules and Boeke, in preparation), as high as found for retroviruses. The mutation frequency, which is strongly influenced by the replication rate and the competitive ability of the mutant genomes relative to their parental wild-types (Domingo and Holland, 1994), is nevertheless not known for retrotransposons.

The tobacco Tnt1 retrotransposon, a member of the Ty1 family of retroelements, is present in >100 copies in the genome of this plant (Grandbastien *et al.*, 1989). Three copies of this element were cloned after insertion in a nitrate reductase gene (Grandbastien *et al.*, 1989; Grappin, personal communication). The tobacco Tnt1 element is one of the very few plant retrotransposons that has been shown to be transcriptionally active (Pouteau *et al.*, 1991; Grandbastien, 1992). Its expression is strongly regulated and its RNA is only detectable in a few tissues and under certain conditions. It is found at a low level in roots and is absent from the leaves of the plant (Pouteau *et al.*, 1991). Tnt1 RNA accumulates at a high level during the preparation of leaf-derived protoplasts due to the use of fungal hydrolases to degrade the cell wall (Pouteau *et al.*, 1991), and after induction with pathogen elicitors (Pouteau *et al.*, 1994).

We were interested in determining if the tobacco genome contains multiple transcriptionally competent Tnt1 elements and if these elements are transcribed at the same time and in the same tissues. We have used a population-based approach in order to define the nature of the expressed Tnt1 elements and to answer the question of whether the Tnt1 RNA is defined by a unique nucleotide sequence or whether it is composed of a population of closely related but different sequences in a manner analogous to the quasispecies distribution of RNA viruses.

Results

Tnt1 RNA is not a unique sequence

We have analysed the variability within the Tnt1 RNA by cloning and sequencing RT-PCR products obtained from protoplast RNA. The sequence analysed corresponds to the last 372 bp of the RNA molecule and contains 50 bp of the 3′ end of the reverse transcriptase coding region (here called RT), 50 bp of the 3′ untranslated region between the ORF and the 3′ LTR (here called linker), the first 200 bp of the U3 region of the LTR known to contain most of the elements controlling Tnt1 expression (Casacuberta and Grandbastien, 1993) (here called box), and the last 72 bp of the RNA molecule containing the rest of the U3 region and the R region of the 3′ LTR (here called end). The analysis of two sets of 25 sequences obtained from two independent RT-PCR amplifications shows that Tnt1 RNA is not a unique sequence but a population of different but closely related sequences. As shown in Figure 1, only 20% of the sequences of RT-PCR 1 and 28% of the sequences of RT-PCR 2 are identical, the rest being related to these major sequences mostly by point mutations. The most represented sequence appears to be the same in the two independent RT-PCR experiments. The second most represented sequences in both independent amplifications are also identical, and the third sequence represented more than once in RT-PCR 2 is also found in RT-PCR 1. The internal variability ranges from 0 to 2.4% substitutions in RT-PCR 1, and from 0 to 2.1% substitutions in RT-PCR 2, most of these substitutions being transitions (85% in RT-PCR 1 and 76% in RT-PCR 2). Some mutations are frequently found in the amplified sequences. The mutation frequency in these hypervariable positions is very similar within the two independent amplification experiments (see Table II). The variability is not homogeneously distributed all along the sequence and, as shown in Table I, it is lower in the RT and box regions. The sequences of the BI and BII boxes, which are transcriptional activators that probably control Tnt1 transcription (Casacuberta and Grandbastien, 1993), are well conserved in the sequences analysed. In addition, the box region does not contain any of the seven hypervariable positions found in the other three regions, and the mutations found in the RT region lead to conservative amino acid substitutions or are silent. Only one mutation in RT-PCR 1 (A→G, position 16, leading to a Glu→Gly substitution) leads to a non-conservative substitution. Deletions are rare but we have found two sequences in RT-PCR 1 and three in RT-PCR 2 that contain a deletion that exactly coincides with a tandemly repeated BII box. BII is found repeated four times in the rest of the sequences (representing 92% of the protoplast-expressed Tnt1 population in RT-PCR 1 and 88% in RT-PCR 2). No sequences with less than three BII boxes were found in the protoplast-expressed Tnt1 population.

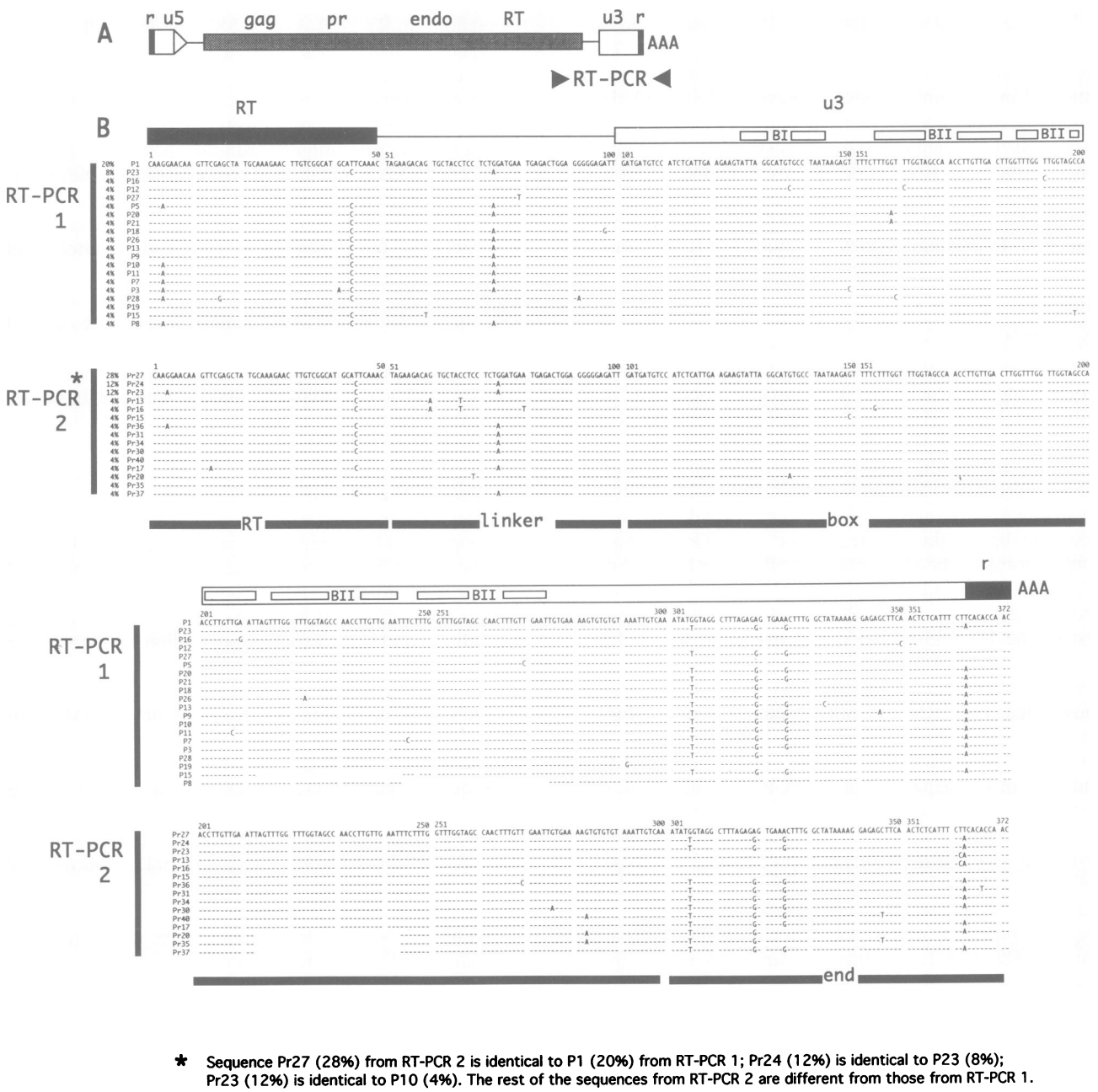
The populations of sequences amplified from the two independent RT-PCR experiments are very similar, suggesting that the variability found reflects the heterogeneity of the Tnt1 RNA population in protoplast RNA. Nevertheless, some of the mutations could be introduced during the amplification procedure. In addition, amplifications of heterogeneous target sequences can lead to artefactual recombinant molecules (Meyerhans *et al.*, 1990). In order to evaluate the background level of mutations due to the PCR amplification, two of the amplified sequences cloned in tandem were subjected to PCR amplification, and 10 amplified sequences were analysed. None of the sequences analysed was a recombinant product, and we have not detected any nucleotide misincorporation within the 3720 nucleotides sequenced (data not shown).

Analysis of the genomic Tnt1 population

We have analysed the variability within the total genomic Tnt1 population by cloning and sequencing 25 Tnt1 partial sequences amplified from tobacco genomic DNA and corresponding to the same regions analysed after amplification from protoplast RNA. As shown in Figure 2, the variability found in the genomic population is much higher than in the protoplast-expressed population. Only two sequences out of 25 are identical (D26 and D29). Two of the sequences differ greatly from the others (D10 and D14). The variability found in the remaining sequences ranges from 0.75 to 4.25% substitutions with respect to the sequence named D2, which is the closest to the consensus sequence. In contrast to what is found for the protoplast-expressed population, the variability is higher in the RT and box regions than in the linker and the end regions. The number of substitutions per nucleotide is as high as 0.24 in the RT region and 0.22 in the box region (see Table I). Nevertheless, most of the new mutations found in the RT coding region are neutral, being silent or leading to conservative substitutions, and only two of them lead to non-conservative substitutions (G→A, position 6 leads to a Asn→Asp substitution and G→A, position 27 leads to a Glu→Lys substitution). In addition to point mutations, deletions are also much more frequent within the box region of the total genomic population. These deletions affect one (44% of the sequences) or two BII boxes (8% of the sequences). The two hypervariable sequences (D10 and D14) do not contain any conserved BI or BII sequences. These results indicate that the box region is the most variable region within the total genomic Tnt1 population although it shows the lowest mutation frequency within the protoplast-expressed population.

The sequence variability within the total genomic population correlates with phenotypic variability

We have previously shown that the U3 region of the LTR contains most of the elements controlling Tnt1 transcription (Casacuberta and Grandbastien, 1993). Mutations within this region could thus affect the transcriptional activity of the different Tnt1 elements. In order to test this hypothesis, we have analysed the ability of the 25 cloned genomic U3 sequences to promote transcription. We cloned a *GUS* reporter gene downstream of the DNA



* Sequence Pr27 (28%) from RT-PCR 2 is identical to P1 (20%) from RT-PCR 1; Pr24 (12%) is identical to P23 (8%); Pr23 (12%) is identical to P10 (4%). The rest of the sequences from RT-PCR 2 are different from those from RT-PCR 1.

Fig. 1. Sequence variability within the protoplast-expressed Tnt1 population. (**A**) Diagram of the relative localization of the sequence analysed within the Tnt1 RNA molecule. (**B**) Nucleotide sequences of two sets of 25 independent clones obtained from two independent RT-PCR amplifications done with tobacco protoplast RNA. Sequences have been named P (for those obtained from RT-PCR 1) or Pr (for those resulting from RT-PCR 2). The sequences have been aligned to the master (most frequently represented) sequence. Dashes indicate sequence identity and blanks indicate deletions. Only the nucleotides which differ from the master sequence are shown. Relative frequencies of each sequence are shown on the left. The relative localization of the sequences with respect to the diagram presented in (A), as well as the localization of BI and BII boxes, is shown above the sequences. The names of the different regions of the analysed sequences are shown underneath.

amplified sequences and performed transient expression experiments using a particle gun approach as previously described (Casacuberta and Grandbastien, 1993). As shown in Figure 3, the 25 amplified sequences, corresponding to partial sequences of 25 Tnt1 genomic elements, promote transcription with varying efficiency under the conditions tested. The sequence variability found thus correlates with phenotypic variability. The major difference between the total genomic population and the protoplast-expressed fraction is the prevalence of sequences containing a high number of tandemly repeated BII boxes within the expressed population of elements. As shown in Figure 3, the lowest level of transcriptional activity is obtained with constructs driven by sequences that do not contain any BII boxes and the strongest promoter activity is shown by sequences containing three or four BII boxes (the number of BII boxes found in the protoplast-expressed Tnt1 sequences). The sequence D13, which only contains two BII boxes, promotes a level of transcription similar to that of most of the sequences containing three BII boxes. Sequence D13 has an insertion of five nucleotides that creates the sequence TAGCCAA analogous to the

Table I. Comparison between the variability found within the total and the protoplast-expressed Tnt1 populations. Frequencies of single base pair substitutions

	Mutation frequency (substitutions/nt)			
	RT region	Linker region	Box region	End region
Protoplast-expressed population (RT-PCR 1–RT-PCR 2)	0.08–0.06	0.1–0.1	0.06–0.03	0.1–0.1
Total genomic population	0.24	0.12	0.22	0.18
Ratio total/expressed (RT-PCR 1–RT-PCR 2)	3–4	1.2–1.2	3.6–7.3	1.8–1.8

core of BII, and which could act, to some extent, as a third BII box.

In addition to the deletions of BII boxes, other mutations are frequently found within the genomic Tnt1 sequences analysed. Some of these mutations clearly have an effect on the Tnt1 promoter, as sequences carrying the same number of BII boxes do not promote transcription with the same efficiency under the conditions tested. It is nevertheless very difficult to determine the effect that individual mutations could have on Tnt1 transcription, based on the results presented in Figure 3, as the amplified genomic Tnt1 sequences often differ from each other by more than a single mutation.

A different population of Tnt1 elements is expressed in roots

In addition to the induction of Tnt1 RNA accumulation found during the preparation of leaf-derived protoplasts (or by treatment of the plant with fungal elicitors), Tnt1 RNA is accumulated at low levels in untreated roots (Pouteau *et al.*, 1991). In order to investigate if the same Tnt1 RNA population is present in both cases, we have analysed the variability within the Tnt1 RNA present in root RNA. We have analysed the last 372 bp of the Tnt1 RNA as we did for the protoplast-expressed Tnt1 population. The result of the sequencing of two sets of 25 sequences amplified from two independent RT-PCR experiments is shown in Figure 4. All the amplified sequences are unique within each one of the two independent RT-PCR experiments, although some sequences are found in both amplified populations. Only one sequence (sequence A2 from RT-PCR 1) is identical to the major protoplast-expressed sequence (sequence P1 from RT-PCR 1 or Pr27 from RT-PCR 2, see Figure 1) which represents >20% of the protoplast-expressed population. The rest of the sequences differ from the sequence A2 mostly by point mutations, but single nucleotide insertions and deletions are also found. Sequences A19 and A24 from RT-PCR 1 and Ar27 and Ar46 from RT-PCR 2 are highly divergent and they present a high degree of internal variability that ranges from 8 to 11% substitutions with respect to sequence A2. The mutation frequency is homogeneously distributed all along the sequence for these highly variable sequences. The internal variability for the rest of the sequences ranges from 0 to 3.2% substitutions with respect to the sequence A2 for sequences from RT-PCR 1 and from 0.3 to 4% for sequences from RT-PCR 2. The mutation frequency in the different regions for these sequences is slightly higher than that found in the protoplast-expressed population and, even though the box region is again the least variable region, the variability

Table II. Comparison between the protoplast and the root-expressed populations of Tnt1 elements. Analysis of frequently mutated positions

Substitution (position)	Presence(%) in the expressed populations			
	Protoplast		Root	
	RT-PCR 1	RT-PCR 2	RT-PCR 1	RT-PCR 2
G→A (4)	28	16	20	8
T→C (44)	64	56	76	76
G→A (74)	50	48	64	64
T→G (106)	not present	not present	28	20
G→A (132)	not present	not present	28	20
G→A (147)	not present	not present	24	24
C→T (234)	not present	not present	28	36
G→A (283)	not present	12	24	12
G→T (305)	60	60	65	83
A→G (319)	60	60	56	66
A→G (325)	48	52	72	74
T→A (363)	58	60	60	64

The positions of the substitutions are shown with respect to the master sequence expressed in protoplasts (Sequence P1, Figure 1). Mutations found at significantly different frequencies in the protoplast and root-expressed populations are boxed.

seems to be much more homogeneously distributed all along the sequence than in the protoplast-expressed population of sequences. We found 0.12 or 0.14 substitutions/nucleotide (subs/nt) for the RT-PCR 1 or RT-PCR 2 sequences in the RT region (mostly silent at the amino acid level), 0.08 or 0.14 subs/nt in the linker region, 0.1 subs/nt for both experiments in the box region and 0.15 or 0.18 subs/nt in the end region (see Table I for a comparison with the variability of the protoplast-expressed population). As in the case of the protoplast-expressed population of Tnt1 sequences, there are several positions that are frequently mutated. The mutation frequency of these positions is very similar within the two independent amplification experiments (see Table II). Most of these frequently mutated positions coincide with those found in the protoplast-expressed population, even though the mutation frequency is in general higher for the root-expressed population. Nevertheless, four new hypervariable positions (shown by closed arrows in Figure 4) appear in the sequences amplified from root RNA. These root-specific frequently mutated positions appear with a similar frequency within the two independent amplification experiments (see Table II) and are all located within the box region. As in the case of the protoplast-transcribed population, most of the sequences found in the root RNA contain four BII boxes. Nevertheless, the number of sequences containing deletions of BII boxes is higher in

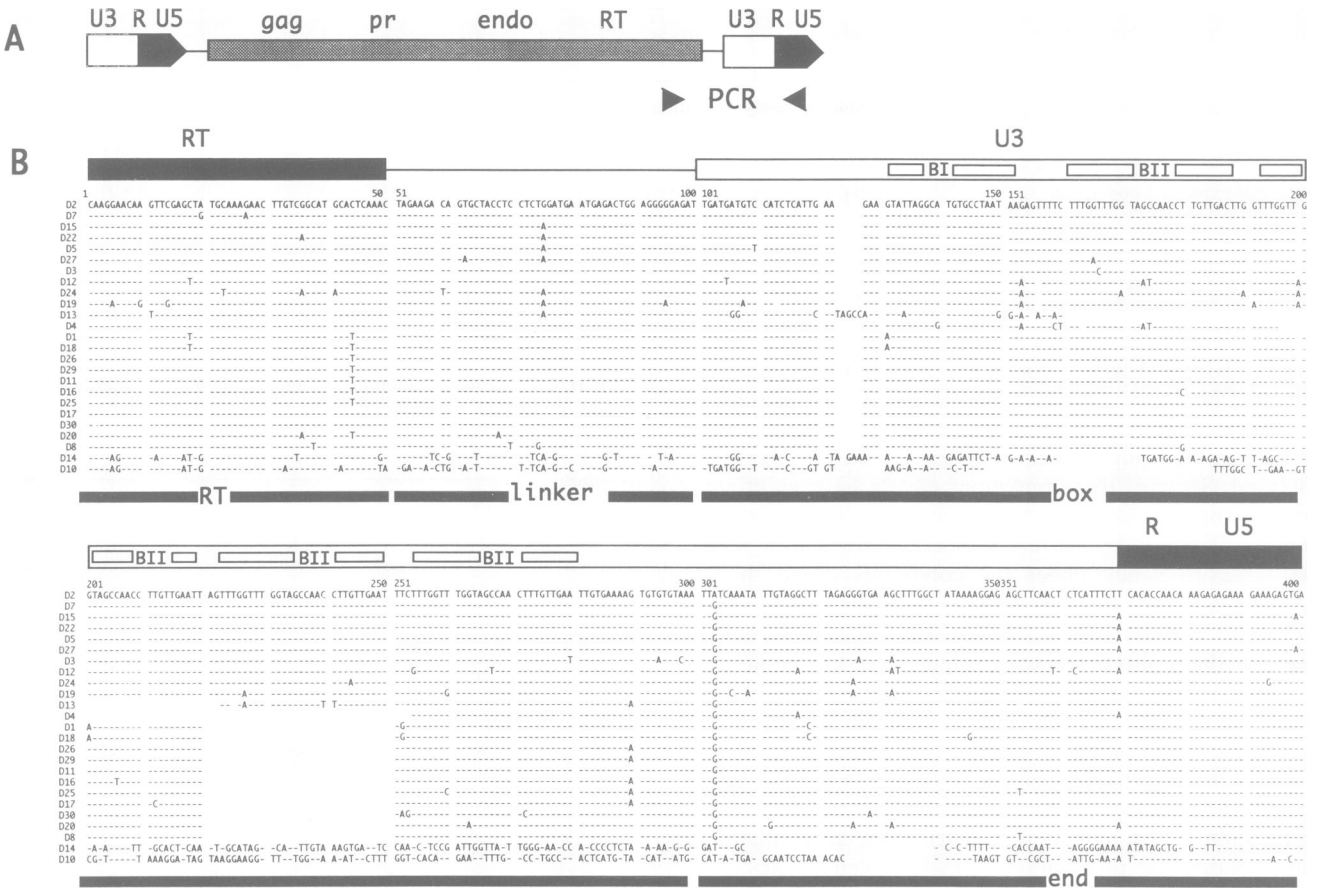


Fig. 2. Sequence variability within the total Tnt1 genomic population. (**A**) Diagram of the relative localization of the sequence analysed within the Tnt1 DNA molecule. (**B**) Nucleotide sequences of 25 independent clones amplified from tobacco genomic DNA. Sequences have been aligned with respect to the D2 sequence. Dashes indicate sequence identity and blanks indicate deletions. Only the nucleotides which differ from the master sequence are shown. The relative localization of the sequences with respect to the diagram presented in (A), as well as the localization of BI and BII boxes, is shown above the sequences. The names of the different regions of the analysed sequences are shown underneath.

the root-expressed population. We found five sequences from RT-PCR 1 (20% of the population) and six sequences from RT-PCR 2 (24%) that contain a deletion of a tandemly repeated BII sequence. A deletion of the BI box is found in the most divergent sequences, A19 and A24 from RT-PCR 1 and Ar27 and Ar46 from RT-PCR 2. In addition, the region deleted in Ar46 also comprises two BII boxes. This sequence is the only sequence amplified from an expressed Tnt1 population that contains fewer than three BII boxes. These results show that the population of Tnt1 elements expressed in roots is different from the one expressed in leaf-derived protoplasts.

When the 100 Tnt1 partial sequences amplified from root or protoplast RNA were compared using the PileUp computer programme of the sequence analysis software package from the Genetics Computer Group Inc. (GCG package), the dendogram obtained clearly reflects the differences between these two populations of expressed elements (Figure 5). Protoplast-amplified sequences tend to plot grouped in the centre of the dendogram while root-expressed sequences usually plot at the ends due to their high degree of divergence. On the other hand, sequences obtained from the two independent RT-PCR amplifications from protoplast RNA are completely mixed in the dendogram, as are those resulting from the two independent RT-PCR assays done with root RNA. This confirms that the differences found between the sequences amplified from

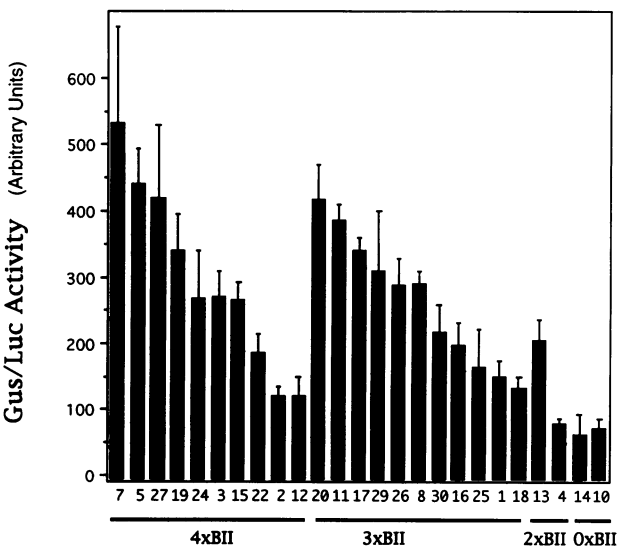
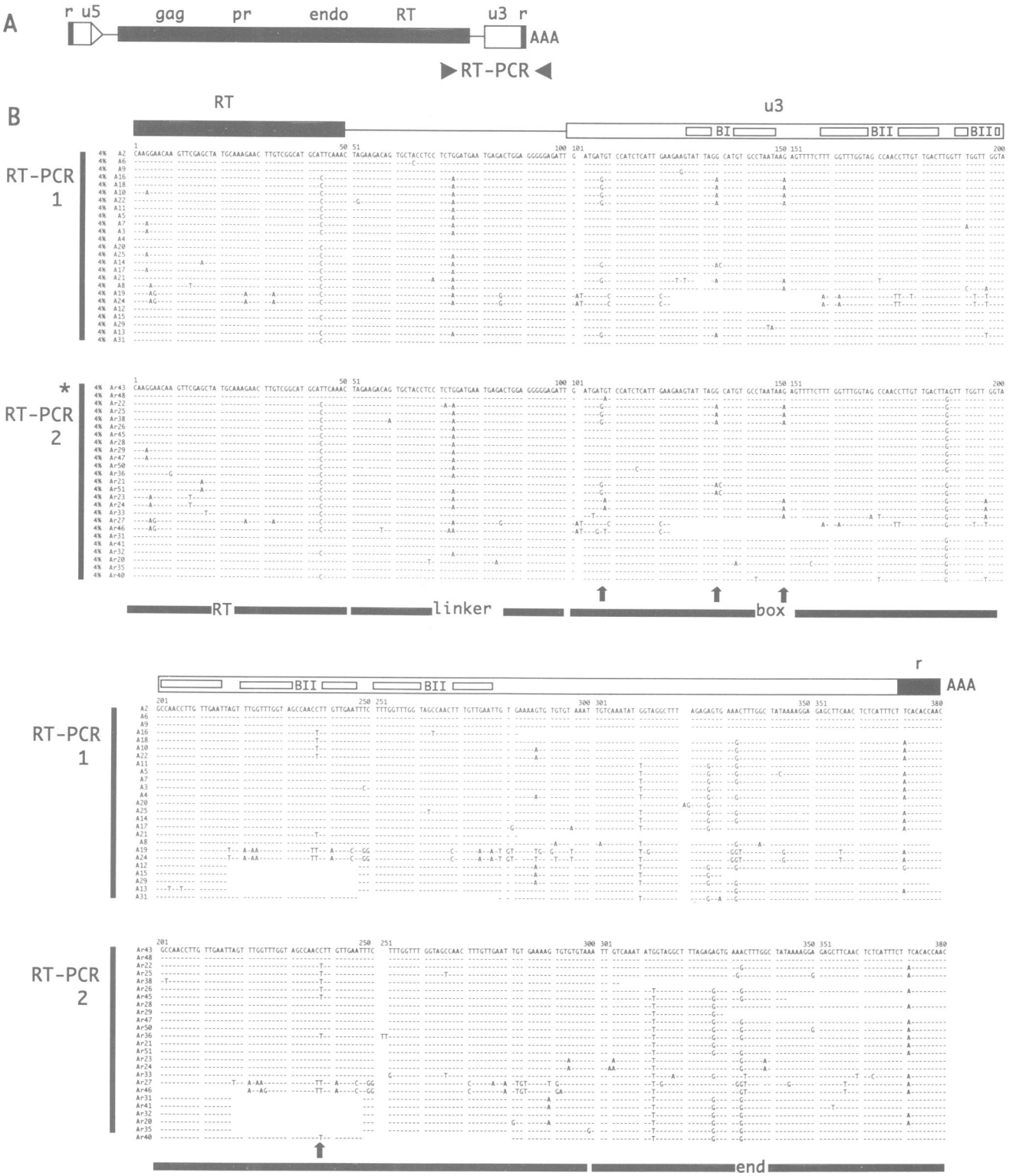


Fig. 3. Promoter activity of the different genomic Tnt1 sequences. Transient expression analysis performed with the reporter gene *GUS* under the control of the different amplified Tnt1 genomic sequences. *GUS* activities in protoplasts are shown relative to an internal luciferase standard. Means of 2–16 independent experiments are shown with their standard errors.



* Sequence Ar45 from RT-PCR 2 is identical to A11 from RT-PCR 1; Ar25 is identical to A16; Ar51 is identical to A14; Ar27 is identical to A24. The rest of the sequences from RT-PCR 2 are different from those from RT-PCR 1.

Fig. 4. Sequence variability within the root-expressed Tnt1 population. (**A**) Diagram of the relative localization of the sequence analysed within the Tnt1 RNA molecule. (**B**) Nucleotide sequences of two sets of 25 independent clones obtained from two independent RT-PCR amplifications done with tobacco root RNA. Sequences have been named A (for those obtained from RT-PCR 1) or Ar (for those resulting from RT-PCR 2). Sequences have been aligned to the sequence most similar to the consensus sequence. Dashes indicate sequence identity and blanks indicate deletions. Only the nucleotides which differ from the top sequence are shown. The relative localization of the sequences with respect to the diagram presented in (A), as well as the localization of BI and BII boxes, is shown above the sequences. The names of the different regions dividing the analysed sequences are shown underneath. Closed arrows indicate frequently mutated positions not present in the protoplast-expressed population.

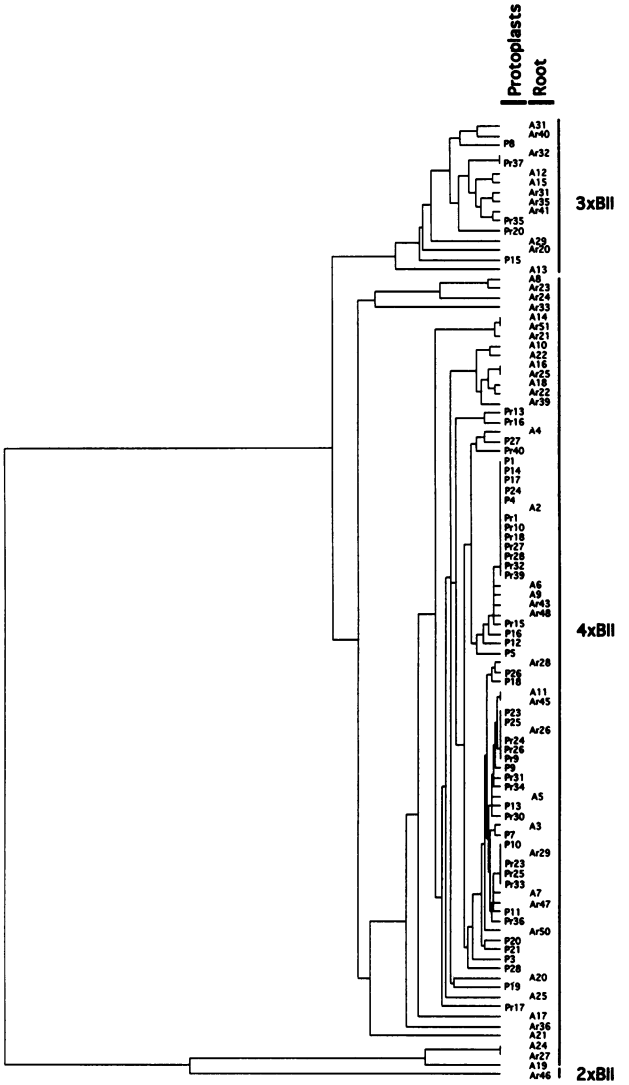


Fig. 5. Comparison between root and protoplast-expressed Tnt1 populations. Dendogram based on similarity scores between every possible pairing of the 100 Tnt1 partial sequences amplified from tobacco RNA. The sequence comparison was done using the PileUp computer programme of the sequence analysis software package from the Genetics Computer Group Inc. (GCG package) which uses the algorithm of Needleman and Wunsch (1970). Sequences obtained from the two independent RT-PCR amplifications from protoplast RNA are shown by P (for those obtained from RT-PCR 1) or Pr (for those resulting from RT-PCR 2). Sequences obtained from the two independent RT-PCR amplifications from root RNA are shown by A (for those obtained from RT-PCR 1) or Ar (for those resulting from RT-PCR 2).

root and protoplast RNA are not due to PCR artefacts, but reflect that the Tnt1 RNA population present in tobacco roots is indeed different from the one found in leaf-derived protoplasts.

Discussion

Quasispecies-like nature of Tnt1 RNA

The analysis of Tnt1 RNAs present in protoplast and root show that when Tnt1 is transcribed its RNA is not a unique sequence but a population of different but closely related sequences. The variability found is very high. All the sequences amplified from root RNA are different within a single amplification experiment, and <30% of the sequences amplified from protoplast RNA are identical within the 372 bp region analysed. If a similar level of variability occurs within the rest of the 5.2 kb Tnt1 RNA, it might be very difficult to find two identical Tnt1 RNA molecules. This highly heterogeneous population of Tnt1 RNAs closely resembles the quasispecies distribution of retrovirus and RNA virus populations. The different sequences in a viral quasispecies are in a rapid equilibrium (Domingo, 1989) due to high mutation rates coupled with extremely rapid replication cycles. Retrotransposition of Tnt1 elements seems to be a rare and very controlled event and only the new elements that integrate into the genome in the germline cells will be transmitted to the descendants. The different Tnt1 RNA sequences are thus not in a rapid equilibrium and they do not behave as a viral quasispecies. The highly variable population of Tnt1 RNAs constitutes a different type of structure that we can define as 'quasispecies-like'. The quasispecies-like nature of Tnt1 RNA is a consequence of both the high number of mutations introduced during the retrotransposition process and the high copy number of Tnt1 elements that the tobacco genome can support. Random genomic mutations can also contribute to some extent to the high sequence variability of the Tnt1 population. Nevertheless, the high mutation rate associated with retrotransposition implies that, for active Tnt1 copies, many mutations will be introduced during this process. In addition, the type of mutations that we have found are similar to those found in retroviral populations, and are, as discussed below, consistent with mutations introduced during error-prone reverse transcription. We thus think that error-prone reverse transcription is probably the major source of the sequence variability found in the Tnt1 population.

This highly heterogeneous nature of the expressed Tnt1 elements implies that the analysis of a particular copy of the element will not necessarily reflect the characteristics of the whole population.

Frequent mutations in the box region lead to inactivation

It is interesting to note that the variability found is not homogeneously distributed all along the sequence analysed. The box region presents the lowest mutation frequency within the transcribed elements, whereas it is the most variable region within the total genomic population. These results indicate that this region contains elements whose sequence and structure are essential for the expression of a given Tnt1 copy and, at the same time, that mutations occur more often within these elements than in the rest of the sequence. This is particularly clear for the tandemly repeated BII element, a transcriptional activator that probably controls Tnt1 expression (Casacuberta and Grandbastien, 1993). Whereas only 40% of the Tnt1 elements contain four tandemly repeated BII boxes, >90% of the protoplast-expressed elements (and 85% of the root-expressed elements) contain four of these repeated elements. We have not found any Tnt1 sequence in protoplast RNA containing less than three BII boxes (and only sequence Ar46 contains two BII boxes within the 50 sequences amplified from root RNA), although these types of sequences represent 16% of the total genomic Tnt1 population. In three Tnt1 elements cloned after insertion

in a tobacco nitrate reductase gene, the BII box was found repeated three (Tnt1–94) (Grandbastien *et al.*, 1989) or four times (Tnt1–95 and Tnt1–96) (Grappin, personal communication) in tandem. This heterogeneity in the number of BII boxes could be caused by deletions or duplications of such elements. The four BII elements located in the U3 region of Tnt1 are not identical, containing a few base differences which allow them to be distinguished from one another. All the sequences analysed here present the same order of BII boxes and no repetitions of a given BII box are found in these sequences. This seems to indicate that the variant sequences arise from a four BII box-containing sequence and that deletions are much more frequent than duplications. It has been reported that large direct repeats are unstable in retroviruses (Rhode *et al.*, 1987; Varela-Echevarria *et al.*, 1993), and that deletion of repeated sequences frequently occurs during Ty1 retrotransposition (Xu and Boeke, 1987). In retroviruses, deletions of tandemly repeated sequences seem to be much more frequent than duplications of such sequences and could be caused by misalignment during reverse transcription (Pathak and Temin, 1990). The apparent low affinity of the enzyme for the template and the presence of two RNAs in the VLP (viral particles for retroviruses) facilitate these types of mutations. In addition, it has been reported that local secondary structures such as stem-loops could facilitate deletion mutations during retroviral minus strand DNA synthesis (Pathak and Temin, 1992). Computer predictions have shown that the tandemly repeated BII sequences could show strong RNA secondary structures (data not shown). Deletion of BII boxes could thus be a frequent event during Tnt1 transposition, leading to the inactivation of some of the newly transposed copies.

The U3 region of retroviruses and retrotransposons is present in both the 5′ and 3′ LTRs, promoting transcription but also determining transcription termination. The U3 region is also found in the 3′ region of the RNA copy of the element. Thus the inhibitory effect of deleting BII boxes could be explained by a loss of transcriptional activity, a defect of transcription termination or a decrease in mRNA stability. We have previously shown that BI and BII boxes are transcriptional activators (Casacuberta and Grandbastien, 1993). The transient expression analysis presented in Figure 3 shows that most of the U3 regions analysed containing three or four tandemly repeated BII boxes promote transcription more efficiently than the U3 regions analysed that contain less than three BII elements under our conditions. The deletion of BII boxes could thus result in a loss of transcriptional activity. However, the relationship found between the number and conservation of BII boxes and the strength of a particular promoter is not clear enough to explain fully the results obtained when comparing the expressed Tnt1 sequences with the total genomic Tnt1 population. As amplified genomic U3 sequences often differ from each other by more than a single mutation, we are currently investigating the effect that single mutations (deletions and point mutations) introduced by site-directed mutagenesis have on the promoter activity of a particular U3 sequence, in order to assess this question unambiguously. We are also investigating the role that the number of BII boxes could play in transcription termination efficiency and RNA stability.

Control of Tnt1 activity and population equilibrium

The major difference between retroviruses and retrotransposons is that retrotransposons lack the capacity to be infectious and to colonize new hosts. Thus retrotransposons are restricted to their host and are transmitted vertically. This implies that, as retrotransposition is a highly mutagenic event, only the existence of strong control mechanisms can allow active elements to subsist within viable genomes. Tnt1 transposition seems to be controlled mainly at the expression level as its RNA accumulates only under very specific conditions (Grandbastien *et al.*, 1994). We show here that mutations (deletions in particular) are especially frequent in elements that are important for Tnt1 expression, and thus the generation of defective Tnt1 copies should be a frequent event. The structure of regulatory elements in tandemly repeated sequences, that are unstable during retrotransposition, could help to maintain the number of active Tnt1 elements under a threshold number in order to avoid affecting tobacco viability. These defective copies arising from the frequent mutagenic events could also compete with the active ones for the factors needed for their expression. The retrotransposon strategy will thus be similar to that of viruses in persistent infections in which the production of mutants, including defective interfering viruses, could help to maintain the non-lytic phenotype (Holland, 1991; Oldstone, 1991). An informative way in which to picture retrotransposon populations like Tnt1 could thus be as extremely persistent viral infections.

A high mutation rate associated with Tnt1 transposition does not however necessarily imply that a Tnt1 population will be continuously evolving at a high rate. When selective conditions remain unchanged, viral quasispecies can reach an equilibrium during which the average or consensus sequence will be maintained for long periods of time, although the mutation rate will be maintained at an invariably high level (Domingo and Holland, 1994). In a similar way, a high mutation rate during Tnt1 transposition could nevertheless be associated with a stable Tnt1 sequence population.

Adaptive capacity of Tnt1 populations

We show here that the population of Tnt1 elements expressed in roots is different from the one expressed in leaf-derived protoplasts. The variability in the root-expressed population is slightly higher than in the protoplast-expressed population. This is especially clear for the region which contains the BI and BII boxes that seem to be essential for the expression of Tnt1 elements in protoplasts. Expression of Tnt1 in roots is much lower than in leaf-derived protoplasts (Pouteau *et al.*, 1991), and the BI and BII elements may be less important for this low level expression. However, BI and BII elements probably do play a certain role in the expression of Tnt1 in roots as, even though the variability of the root-expressed population of elements within this region is higher than that of the protoplast-expressed population, it is still clearly lower than in the total genomic population of Tnt1 elements.

The presence of different populations of Tnt1 expressed under different conditions closely resembles the evolution of viral quasispecies under different selective conditions

(Domingo and Holland, 1994), suggesting an adaptive capacity for the Tnt1 retrotransposon population.

The high homology shared between plant, animal and yeast retrotransposons has led some researchers to propose that horizontal transmission of these elements could be a rare but possible event (Xiong and Eickbush, 1990; Flavell *et al.*, 1992). In the case of Tnt1, the fact that its expression is induced by pathogens might allow this transfer. The great adaptability of viral quasispecies facilitates their exploitation of new ecological niches (Domingo and Holland, 1994) and we propose that the simultaneous presence of different Tnt1 RNA sequences could facilitate the evolution of Tnt1 elements horizontally transferred to a new host into a population capable of co-existing with this new host without compromising its viability, as is the case in tobacco.

Materials and methods

RNA and DNA extractions, PCR amplification, cloning and sequencing

RNAs from protoplasts and roots of *Nicotiana tabacum* (cv Xanthi) wild-type line XHFD8 (Bourgin and Missonier, 1973) were obtained as described previously (Pouteau *et al.*, 1991). DNA were obtained according to Dellaporta *et al.* (1984). Root or protoplast RNA (5 μg) was reverse transcribed into cDNA using standard protocols with an oligo(dT) primer containing a 5′ extension [5′-GCTAGGCCACTGTGGCCT(15)-3′]. Genomic DNA (200 ng) or 1/20 of the cDNA obtained was PCR amplified for 35 cycles (45 s 94°C, 30 s 55°C, 1 min 72°C) using as 5′ primer an oligonucleotide in the region coding for the reverse transcriptase (5′-TATGCTGACCAAGGTGGTAC-3′) and as 3′ primer the same oligo(dT) primer used for the first strand synthesis for the cDNA and an oligo in the U5 region of the LTR (5′-TTATACCTTGTCTGTG-AAACC-3′) for the genomic DNA. PCR-amplified products were cloned in a pUC18 plasmid digested with *Hin*cII in which a single thymidine was added to the 3′ ends according to Marchuk *et al.* (1990). The sequencing of the clones was done with a PRISM sequencing kit (Applied Biosystems) and sequences were analysed in a automatic sequencing apparatus (Applied Biosystems).

GUS constructs and transient expression analysis

pUC18 clones containing sequences amplified from total genomic DNA were digested with *Bam*HI and *Pst*I, filled-in and cloned in the *Bam*HI site of a pG46-derived plasmid (Casacuberta and Grandbastien, 1993) containing the reporter gene β-glucuronidase (*GUS*) (devoid of promoter) and the terminator of the pea rbc S3C gene. Tobacco leaves were bombarded with 1 μg of the different constructs analysed, mixed with 0.5 μg of a plasmid containing a luciferase (*LUC*) reporter gene driven by the 35S promoter of the cauliflower mosaic virus which was used as an internal standard. Protoplasts were prepared from the bombarded leaves in order to activate transcription. *GUS* and *LUC* activity were measured as previously described (Casacuberta and Grandbastien, 1993).

Acknowledgements

We are extremely grateful to J.Gomez for suggesting that we should analyse Tnt1 variability and for very helpful discussions and useful comments on the manuscript. We thank H.Lucas, C.Robaglia and I.Small for critical reading of the manuscript. We also wish to thank A.Gabriel for sharing unpublished data with us. J.M.C. is supported by an INRA post-doctoral fellowship.

References

Boeke,J.D. and Corces,V. (1989) *Annu. Rev. Microbiol.*, **45**, 403–434.
Boeke,J.D., Garfinkel,J., Styles,C.A. and Fink,G.R. (1985) *Cell*, **40**, 491–500.
Boeke,J.D., Eichinger,D., Castrillon,D. and Fink,G.R. (1988) *Mol. Cell. Biol.*, **8**, 1432–1442.
Bourgin,J.P. and Missonier,C. (1973) *Haploid Inform. Serv.*, **8**, 7.
Casacuberta,J.M. and Grandbastien,M.-A. (1993) *Nucleic Acids Res.*, **21**, 2087–2093.
Coffin,J.M. (1986) *Cell*, **46**, 1–4.
Dellaporta,S.L., Wood,J. and Hicks,J.B. (1984) *Molecular Biology of Plants: A Laboratory Course Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Domingo,E. (1989) *Prog. Drug. Res.*, **33**, 93–133.
Domingo,E. and Holland,J.J. (1994) In Morse,S.S. (ed.), *Evolutionary Biology of Viruses.* Raven Press, NY, pp. 161–184.
Domingo,E. *et al.* (1985) *Gene*, **40**, 1–8.
Flavell,A.J., Smith,D.B. and Kumar,A. (1992) *Mol. Gen. Genet.*, **231**, 233–242.
Grandbastien,M.-A. (1992) *Trends Genet.*, **8**, 103–108.
Grandbastien,M.-A., Spielmann,A. and Caboche,M. (1989) *Nature*, **337**, 376–380.
Grandbastien,M.-A., Audeon,C., Casacuberta,J.M., Grappin,P., Lucas,H., Moreau,C. and Pouteau,S. (1994) *Genetica*, **93**, 181–189.
Holland,J.J. (1991) In Fields,B.N. and Knipe,D.N. (eds), *Fundamental Virology.* Raven Press, NY, pp. 151–165.
Hu,W.-S. and Temin,H. (1990a) *Proc. Natl Acad. Sci. USA*, **87**, 1556–1560.
Hu,W.-S. and Temin,H. (1990b) *Science*, **250**, 1127–1233.
Katz,R.A. and Skalka,A.M. (1990) *Annu. Rev. Genet.*, **24**, 409–445.
Marchuk,D., Drumm,M., Saulino,A. and Collins,F.S. (1990) *Nucleic Acids Res.*, **19**, 1154.
Meyerhans,A., Cheynier,R., Albert,J., Seth,M., Kwok,S., Sninsky,J., Morfeldt-Manson,L., Asjo,B. and Wain-Hobson,S. (1989) *Cell*, **58**, 901–910.
Meyerhans,A., Vartanian,J.P. and Wain-Hobson,S. (1990) *Nucleic Acids Res.*, **18**, 1687–1691.
Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453
Oldstone,M.B.A. (1991) *J. Virol.*, **65**, 6381–6386.
Pathak,V.K. and Temin,H.M. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 6024–6028.
Pathak,V.K. and Temin,H.M. (1992) *J. Virol.*, **66**, 3093–3100.
Pouteau,S., Huttner,E., Grandbastien,M.-A. and Caboche,M. (1991) *EMBO J.*, **10**, 1911–1918.
Pouteau,S., Grandbastien,M.-A. and Boccara,M. (1994) *Plant J.*, **5**, 535–542.
Rhode,B.W., Emerman,M. and Temin,H.M. (1987) *J. Virol.*, **61**, 925–927.
Steinhauer,D.A. and Holland,J.J. (1987) *Annu. Rev. Microbiol.*, **41**, 409–433.
Temin,H.M. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 6900–6903.
Varela-Echevarria,A., Prorock,C.M., Ron,Y. and Dougherty,J.P. (1993) *J. Virol.*, **67**, 6357–6364.
Varmus,H. and Brown,P. (1989) In Berg,D.E. and Howe,M.M. (eds), *Mobile DNA.* American Society for Microbiology, Washington DC, pp. 53–108.
Wain-Hobson,S. (1993) *Curr. Opin. Genet. Dev.*, **3**, 878–883.
Xiong,Y. and Eickbush,T.H. (1990) *EMBO J.*, **9**, 3353–3362.
Xu,H. and Boeke,J.D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 8553–8557.

Received on September 7, 1994; revised on February 13, 1995