Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 759–764, January 1996
Plant Biology

Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution

(plant DNA viruses/replication origin/*AL1* gene/illegitimate recombination/horizontal transmission)

EDUARDO R. BEJARANO*, ALAA KHASHOGGI, MICHAEL WITTY, AND CONRAD LICHTENSTEIN†

Department of Biochemistry, Imperial College of Science, Technology and Medicine, Exhibition Road, London SW7 2AZ, United Kingdom

Communicated by Mary-Dell Chilton, Ciba–Geigy Corporation, Research Triangle Park, NC, October 13, 1995 (received for review December 1, 1994)

ABSTRACT **Integration of viral DNA into the host nuclear genome, although not unusual in bacterial and animal systems, has surprisingly not been reported for plants. We have discovered geminivirus-related DNA (GRD) sequences, in the form of distinct sets of multiple direct repeats comprising three related repeat classes, situated in a unique locus in the *Nicotiana tabacum* (tobacco) nuclear genome. The organization of these sequences is similar or identical in eight different tobacco cultivars we have examined. DNA sequence analysis reveals that each repeat has sequences most resembling those of the New World geminiviral DNA replication origin plus the adjacent *AL1* gene, encoding the viral replication protein. We believe these GRD sequences originated quite recently in *Nicotiana* evolution through integration of geminiviral DNA by some combination of the processes of illegitimate recombination, amplification, deletions, and rearrangements. These events must have occurred in plant tissue that was subsequently able to contribute to meristematic tissue yielding gametes. GRD may have been retained in tobacco by selection or by random fixation in a small evolving population. Although we cannot detect transcription of these sequences, this does not exclude the possibility that they may originally have been expressed.**

Nature provides many examples of recombination between viral DNAs and the host genome such as site-specific integration of bacteriophage λ and retroviral integration into quasi-random target sites, requiring viral recombinases, and, independent of such recombinases, "illegitimate" integration of simian virus 40 (SV40) (ref. 1 and references therein). Most plant viruses have RNA genomes and some have been shown to acquire host sequences by RNA recombination with host transcripts (2). Two well-studied plant DNA viruses are the double-stranded (ds) circular caulimoviruses and the single-stranded (ss) circular geminiviruses. Caulimoviruses replicate via an RNA intermediate by using a virus-encoded reverse transcriptase, but no associated integrase activity has been identified (3). Geminiviruses have genomes of one or two components and replicate via a circular dsDNA intermediate (ref. 4 and references therein). Curiously, although these DNA viruses replicate in the nucleus, there are no reports of their recombination with the plant nuclear genome. Here we present our discovery of multiple repeats of a segment of geminiviral DNA in the nuclear genome of *Nicotiana tabacum* (tobacco).

MATERIALS AND METHODS

General Methods. All general recombinant DNA methods, buffer compositions, bacterial strains used, and references to original publication of methods are as described by Sambrook *et al.* (5).

Southern Blot Analysis of Tobacco DNA. Tobacco DNA was prepared essentially according to Dellaporta *et al.* (6). Tobacco DNA was digested with a variety of restriction endonucleases and 10 μg was fractionated through a 0.8% agarose gel in Tris/borate/EDTA buffer, transferred to Hybond N nylon membranes (Amersham), and hybridized with ^{32}P-radiolabeled probes in Church buffer (5) overnight at 65°C. The final two washes of the membranes were in 0.2× standard saline citrate (SSC) at 60°C and labeled bands were visualized by autoradiography. The probes used were as follows, with nucleotide numbers reading clockwise in Fig. 1*A*: the *AL1Δ2/3* region and the 5′, middle, and 3′ probes encompassing nucleotide positions 1358–13, 2059–18 (from an *Nco* I site to 6 bases from the ATG initiation codon of *AL1*), 1762–2058 (an *Apa* I–*Nco* I fragment) and 1356–1761 (*Bam*HI–*Apa* I fragment), respectively. The AR1 probe was a *Sau* I fragment (nt 1298–104). All nucleotide depositions are as defined in the EMBL accession sequence K02029.

Construction of Genomic Library of Tobacco DNA. Tobacco DNA was partially digested with *Sau*3AI, ligated into *Bam*HI-digested EMBL4 λ arms, and packaged *in vitro* with Gigapack II Gold packaging extract. Recombinant clones (*red*$^-$, *gam*$^-$) were selected by *spi*$^-$ selection on *Escherichia coli* strain Q359 [*supE, hsdR*, (P2)]. Following plaque screening of 5×10^5 recombinant clones probed with *AL1Δ2/3* 10 positive plaques were identified and amplified in *recA*$^-$ hosts (*E. coli* XL1-Blue or DH5α).

Restriction Mapping of Recombinant λ Clones. Clones were linearized by digestion with *Xma* I, end-labeled by filling in with [α-^{32}P]dCTP, and digested with *Xba* I to release the label at one end. Partial digestions were then performed with *Bam*HI. The series of partial digestion products was fractionated through a 0.8% agarose gel. The gel was dried and bands were visualized by autoradiography.

DNA Sequence Analysis. This was performed by the Sanger chain-termination method. In all cases both strands were sequenced. λ clones were subcloned into M13 or plasmid vectors and sequence was extended both by use of exonuclease III-generated nested deletions and by use of internal primers synthesized on a Pharmacia GeneAssemblerPlus.

Contour-Clamped Homogeneous Electric Field (CHEF) Gel Analyses. Tobacco protoplasts, isolated as described by Draper *et al.* (7), were suspended at 10^6 cells per ml in 1% low-melting-point agarose (IBI; a grade melting at 65°C)/0.7 M manni-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: GRD, geminivirus-related DNA; TGMV, tomato golden mosaic virus; ORF, open reading frame; SV40, simian virus 40; ds, double-stranded; SS, single-stranded.

*Present address: Unidad de Genética, Departamento de Biología Celular y Genética, Facultad de Ciencias, Campus Universitario de Teatinos 29017, Málaga, Spain.

†Present address: School of Biological Sciences, Queen Mary and Westfield College, Mile End Road, London E1 4NS, United Kingdom.

tol/10 mM $CaCl_2$, pH 5.8. The molten agar was placed in 2 mm × 5 mm × 10 mm moulds and allowed to solidify for 1 hr at 4°C. Blocks were incubated for 48 hr at 50°C in lysis buffer (0.5 M EDTA, pH 8.0/1% sodium *N*-lauroylsarcosine/0.2% proteinase K). Prior to digestion the blocks were washed twice at 50°C and ten times at 20°C, 30 min per wash in TE (10 mM Tris/10 mM EDTA, pH 8.0) plus 1 mM phenylmethanesulfonyl fluoride. After two incubations in 1 ml of restriction enzyme buffer plus 8 mM spermidine for 1 hr at 20°C, DNA was digested overnight with *Sal* I at 37°C in 250 μl of fresh buffer. Digestion was terminated by incubation with 1 ml of TE plus 0.1% proteinase K. The digested DNA samples were run on CHEF gels (1% agarose, 1× Tris/borate/EDTA buffer) as described by Chu *et al.* (8). After Southern blotting, filters were probed with ^{32}P-radiolabeled GRD3-1 and GRD5 probes, and bands were visualized by autoradiography.

Estimation of Copy Number of Genomic Geminivirus-Related DNA (GRD). Assuming a haploid genome size of 4.4 × 10^9 bp (9), 1 μg of *N. tabacum* total DNA contains 2.27 × 10^5 haploid genomic copies of DNA (neglecting the contribution of chloroplastic and mitochondrial DNAs); 0.2 pg of the ≈900-bp GRD3-3 sequence contains an equivalent number of molecules. Two methods were used to estimate the copy number of GRD: slot blots and Southern blots. For slot blots, samples (0.2, 1, and 2 μg) of total tobacco (cv. Samsun) DNA were spotted onto a membrane in a slot-blot apparatus adjacent to 0, 0.2, 1, 2, 10, 20, 100, and 200 pg of GRD3-3 mixed with 1 μg of salmon sperm DNA. Southern blot analysis was performed as described above, except that total tobacco DNA was digested with *Bam*HI and samples of 0.01, 0.1, 0.5, and 1 μg were loaded in the same gel as 0, 2, 20, 100, and 200 ng of GRD3-3 mixed with 1 μg of salmon sperm DNA. Both filters were probed with ^{32}P-radiolabeled GRD3-3. The final wash was under high-stringency conditions (0.1× SSC/0.1% SDS, 65°C). The copy number of genomic GRD repeats was estimated with a PhosphorImager (Molecular Dynamics) where, for the Southern blot, data for all bands were added together. Both experiments were done twice.

RESULTS

Tobacco Genomic Sequences Cross-Hybridize with Tomato Golden Mosaic Virus (TGMV) DNA. The two genomic components, TGMVA (Fig. 1*A*) and TGMVB, of TGMV, a typical bipartite geminivirus, share similar 235-base intergenic regions which carry all cis-essential replication sequences including a DNA binding site for AL1, a trans-acting, TGMVA-encoded protein essential for replication of both genome components (ref. 10; see also ref. 11 and references therein). We previously constructed TGMV-resistant *N. tabacum* (tobacco) plants expressing a viral antisense RNA (12) and serendipitously discovered that genomic DNAs of untransformed controls cross-hybridized with a TGMV probe encompassing the *AL1* coding sequence and the 5′ ends of the *AL2* (overlapping) and *AL3* genes (designated *AL1Δ2/3*). These GRD sequences are shown in Fig. 1*B*, which is a representative genomic Southern blot of two tobacco cultivars, Samsun and Petite-Havana SR1. Such cross-hybridization is seen only with probes specific for the 5′ and 3′ ends of *AL1*, but not with probes for *AL2*, *AL3*, the middle of *AL1*, or the coat protein gene, *AR1* (data not shown). In further analyses, we found no restriction enzyme-site polymorphism in these cross-hybridizing sequences with these enzymes in a total of eight tobacco cultivars we examined: SR1, Samsun, NC95, SCR1, Havana, White Burley, Speight G28, and Ngana-Ngana (A. Warry, E.R.B., M.W., and C.L., unpublished work).

DNA Sequence Analysis of Cross-Hybridizing Genomic Sequences. The cross-hybridization we observed might result from fortuitous base pairing of the TGMV probe to genomic tobacco DNA or might reflect a common ancestry of these

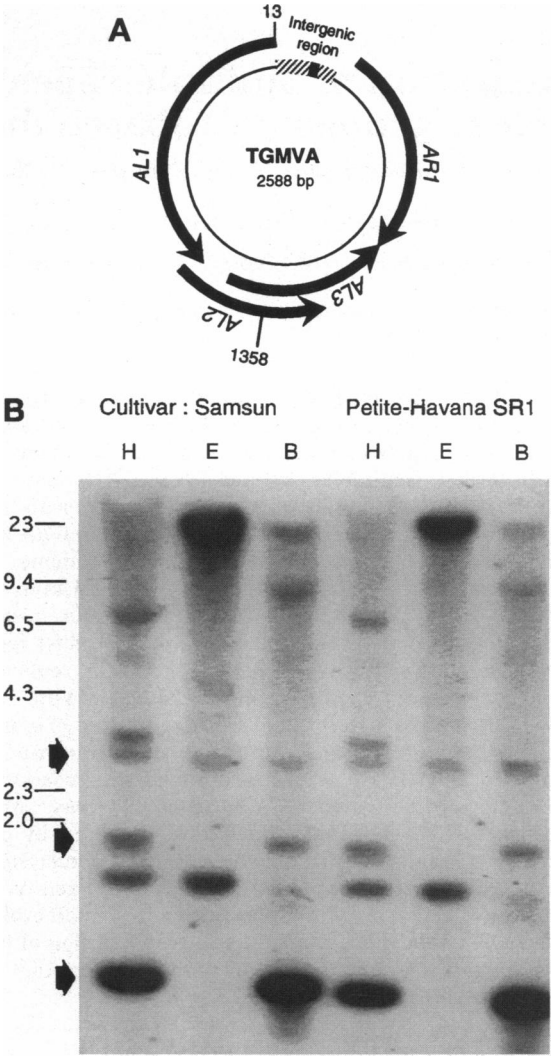

FIG. 1. (*A*) Genetic map of TGMVA (EMBL accession no. K02029) showing four open reading frames (ORFs) (arrows). Only the coat protein gene, *AR1*, is encoded by the viral plus strand. The 235-base intergenic region (hatched box) is similar to that in TGMVB and has a palindromic sequence (in black), capable of forming a stem–loop structure, the putative AL1 protein cleavage site. (*B*) *AL1Δ2/3* probe cross-hybridization with total genomic DNA in Southern blots of *N. tabacum*, cv. Samsun and cv. Petite-Havana SR1. Genomic DNA was digested with *Hin*dIII (H), *Eco*RI (E), or *Bam*HI (B). Positions of molecular size (kb) markers are indicated at left. Note the probable single, double, and triple unit repeats, where some repeats lack the relevant restriction enzyme sites—e.g., for the *Hin*dIII digest, the GRD3 monomer (0.9 kb), dimer (1.8 kb), and trimer (2.7 kb) are indicated by arrows from bottom to top respectively.

sequences. To address this, we constructed a λ library of total genomic Samsun DNA and identified 10 clones hybridizing to probes from the 5′, 3′, or both ends of *AL1*. Restriction enzyme mapping of 4 selected λ clones and DNA sequence analysis showed GRD sequences arranged as multiple contiguous direct repeats (Fig. 2*A*) of three highly related monomer types: GRD5, GRD3, and GRD53 (Fig. 2*B*). Moreover, Southern blot analysis of the 4 λ clones showed that they represent the genomic organization of GRD and have not undergone rearrangements of the repetitive elements during cloning (data not shown). DNA slot blot and Southern blot analyses indicated that there are ≈360 GRD repeats in the tobacco genome (data not shown). CHEF electrophoresis analysis of tobacco protoplast DNA showed a unique ≈340-kb *Sal* I restriction fragment of nuclear DNA that hybridized with a GRD probe, indicating that most if not all GRD repeats are present on this fragment

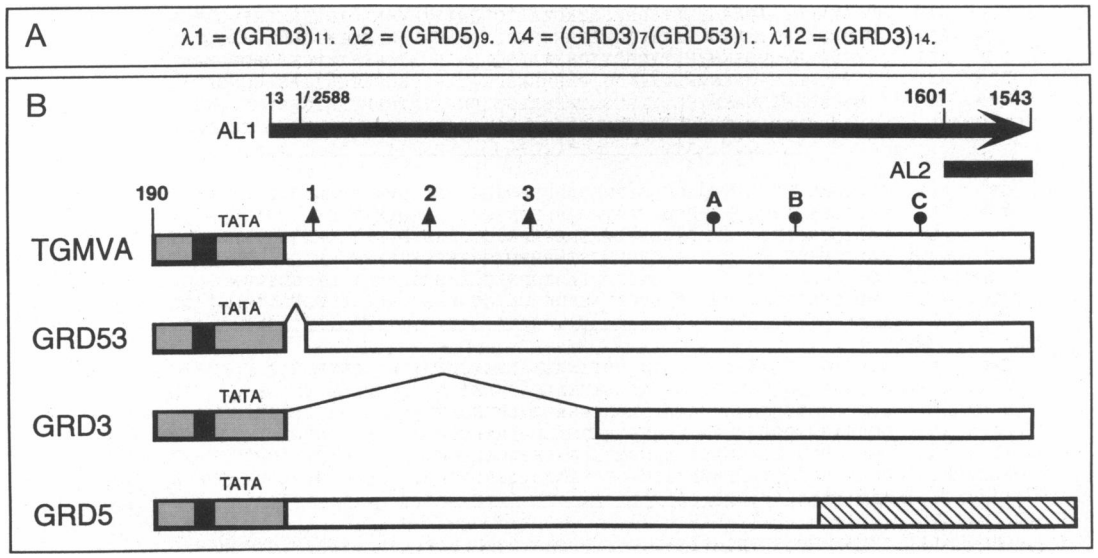

FIG. 2. Diagram of GRD sequences. (*A*) Organization of multiple tandem GRD repeats in four EMBL4 tobacco genomic 1 clones: λ1 carries 11 direct repeats of GRD3 (each GRD3 repeat is 0.9 kb); λ2 carries 9 direct repeats of GRD5 (each GRD5 repeat is 1.4 kb), λ4 carries 7 direct repeats of GRD3 plus one copy of GRD53; λ12 carries 14 direct repeats of GRD3. (*B*) Comparison of GRD5, GRD3, and GRD53 with the *AL1* and *AL2* regions of TGMVA (nucleotide positions as for EMBL K02029), showing regions of GRD repeats similar to geminiviral AL1 coding sequences (open boxes), and intergenic region (stippled boxes) with sequence stem and loop (black box). Also shown is the GRD5 nongeminiviral sequence (hatched box). The vertical arrows numbered 1–3 indicate that GRD retains conserved domains present in rolling circle replication proteins, including AL1 proteins (13, 14). Similarly, black circles A–C indicate the conserved motifs of NTP-binding proteins, including AL1 proteins (15).

(data not shown). This result suggests that a single chromosomal locus encompasses GRD repeats. The same model is also suggested by *in situ* hybridization studies of fluorescence-labeled probes to tobacco chromosomes at mitosis and at diakinesis of meiosis, where GRD repeats were detected at a single site on a small submetacentric pair of homologues (16).

We determined the DNA sequence of GRD53, one copy of GRD5 and six copies of GRD3 (unpublished work). Comparison (Figs. 2*B* and 3) of the DNA sequences and deduced amino acid sequences (where appropriate; see below) of the GRD repeats to those of geminiviruses strongly suggests a common ancestry for GRD and geminiviral DNA, rather than spurious similarity. The lack of strong DNA sequence similarity between the middle region of *AL1* and the genomic copies of GRD explains why we were unable to detect cross-hybridization in the genomic Southern blots discussed above. The multiple GRD repeats have incomplete *AL1* ORFs with no consensus intron/exon splice sites that could yield spliced exons with a more convincing *AL1* ORF, suggesting that none of the GRDs sequenced so far encodes a functional gene product. Further, Northern blot and reverse transcription–PCR analyses of total RNA isolated from root, stem, leaf, and flower sources yielded no evidence of mRNA or antisense RNA transcription (shown with strand-specific sense and antisense *AL1* probes and a dsDNA GRD53 probe). However, we cannot rule out either the possibility of low expression, in specialized cells, over a short window of time during development or that there is a functional, single/low-copy GRD-related (but non-cross-hybridizing) gene that is expressed.

Examination of the potential coding capacity of the GRD repeats indicates that GRD5, GRD3, and GRD53 (the largest iteration) encode amino acid sequences similar to those encoded by the 5′, 3′, and both ends of *AL1*, respectively. The only significant region of DNA which is not similar to geminiviral DNA is the 3′ end of GRD5. By pooling the sequence data and assuming that three frameshift mutations and a point mutation have occurred, it is possible to make a parsimonious model of an ORF that may represent the sequence of a full-length "ancient" geminiviral AL1 protein (Fig. 3) that is significantly similar to other geminiviral AL1 proteins (Table 1). Because of this, the similarity of some individual GRD repeats to the AL1 protein is underrepresented—some repeats contain regions that are more similar to geminiviruses than the consensus model. Here, the CLUSTAL V program was used to align the ancient geminiviral AL1 amino acid sequence with those of AL1 proteins of selected subgroup III geminiviruses (17): TGMV and bean golden mosaic virus (BGMV) were selected as well-studied examples of New World geminiviruses whereas African cassava mosaic virus (ACMV) and Indian cassava mosaic virus (ICMV) were used as examples of Old World geminiviruses. Tomato yellow leaf curl virus (TYLCV) is a relatively distantly related subgroup III geminivirus.

All GRD repeats contain an *AL1*-gene TATA box and a putative 34-base stem–loop structure—e.g., 5′-GCGTCCATCCGGTAATATTATAACGGATGGACGC-3′ in GRD53 (stem sequences underlined)—that corresponds to sequences found in the intergenic region of many geminiviruses. Short inverted repeats that serve as the recognition sites of the cognate AL1 are also retained. These repeats are virus-specific in function; e.g., the AL1 proteins of TGMV and squash leaf curl virus (SqLCV) will not transactivate one another's replication (10). This fact and the presence of base deletions and insertions discovered in all three classes of repeats presumably account for our inability to transactivate ectopic replication of GRD following TGMV infection. However, we can conclude that the putative intergenic region of GRD carries all of the essential motifs of geminiviral intergenic regions.

The putative stem–loop structure, highly conserved in geminiviruses, is similar to replication origin sequences in the ssDNA bacteriophage φX174 (4). This origin is endonucleolytically cleaved by φX174-encoded gpA protein, where a tyrosine in the active site forms a covalent bond with the 5′ end of the strand-specific nicked DNA during rolling-circle replication (reviewed in ref. 20). Moreover, conserved motifs have been identified in 59 proteins mediating the initiation of rolling-circle replication in a broad range of prokaryotic replicons—including ssDNA plasmids of Gram-positive bacteria, ssDNA bacteriophage, and conjugative

```
GRD  AL1  MPPSKKFRIRIQAKNYFLTYRHSSLTKDEALTQLQNISTPVNKLLIRVFRELHEDGEPHL
TGMV AL1   MPSHPKRFQINAKNYFLTYPQCSLSKEESLSQLQALNTPINKKFIKICRELHEDGQPHL
BGMV AL1    MPPPQRFRVQSKNYFLTYPRCTIPKEEALSQLQKIHTTTNKKFIKVCEERHDNGEPHL
ACMV AL1     MRTPRFRVQAKNVFLTYPNCSIPKEHLLSFIQTLSLPSNPKFIKICRELHQNGEPHL
ICMV AL1    MSPPKRFQINAKNYFLTYPRCSLTKEEALSQIRNFQTPTNPKFIKICRELHENGEPHL
TYLCV AL1   MAQPKRFQINAKHYFLSFPKCSLSKEEALEQLLQLQTPTNKKYIKICRELHEDGQPHL
                 *.....*. **.. .. *.. * . . .  *  *.. * *..*.***

GRD  AL1  HVLIQFEGKYVCTNNRAFDLSSPTRSAHFHPNIQGAKSSSDVKTYVEKDGDFIDFGVFQI
TGMV AL1  HVLIQFEGKYCCQNQRFFDLVSPTRSAHFHPNIQRAKSSSDVKTYIDKDGDTLVWGEFQV
BGMV AL1  HALIQFEGKFICTNKRLFDLVSTTRSAHFHPNIQGAKSSSDVKEYIDKDGVTIEWGQFQV
ACMV AL1  HALIQFEGKITITNNRLFDCVHPSCSTNFHPNIQGAKSSSDVKSYLDKDGDTVEWGQFQI
ICMV AL1  HVLIQFEGKYKCQNQRFFDLVSPTRSAHFHPNIQGAKSSSDVKSYIDKDGDTWRWGTFQI
TYLCV AL1 HMLIQFEGKFNCKNNRFFDLVSPTRSAHFHPNIQGAKSSSDVKSYIDKDGDVLEWGTFQI
          * *******    *.* **    . *. ****** ******** *. ***     * **.

GRD  AL1  DGRSARGGCQSANDSYAKAINSGSTINXLKILXEXQPIDYIRDLDKLRGNLDRHFAPPRQ
TGMV AL1  DGRSARGGCQTSNDAAAEALNASSKEEALQIIREKIPEKYLFQFHNLNSNLDRIFDKTPE
BGMV AL1  DGRSARGGQQSANDSYAKALNADSIESALTILKEEQPKDYVLQNHNIRSNLERIFFKVPE
ACMV AL1  DGRSARGGQQSANDAYAKALNSGSKSEALNVIRELVPKDFVLQFHNLNSNLDRIFQEPPA
ICMV AL1  DGRSARGGQQSANDAYAAALNSGSKSEALKILRELAPRDYLRDFHHISSNLDRIFTKPPP
TYLCV AL1 DGRSARGGQQTANDAYAKAINAGSKSEALDVIKELAPRDYILHFHNINSNLDRVFQVPPA
          ******** *.**. * *.*. *    *  .. *  *  .. . . . .**.* *

GRD  AL1  IFVSKWDPQSFVSQMILSSLGEVFEDPSARPVQNNSSDRPLSLILEGPSRTGKTAWARSL
TGMV AL1  PWLPPFHVSSFTN--VPDEMRQWAENYFGKS-SAARPERPISIIIEGDSRTGKTMWARSL
BGMV AL1  PWVPPFPLSSFVN--IPVVMQDWVDDYFGRG-SAARPERPISIIVEGDSRTGKTMWARAL
ACMV AL1  PYVSPFPCSSFDQ--VPDELEEWVADNV-RD-SAARPWRPNSIVIEGDSRTGKTIWARSL
ICMV AL1  PYENPFPLSSFDR--VPEELDEWFHENV-MG-RA-RPLRPKSIVIEGDSRTGKTMWSRAL
TYLCV AL1 PYVSPFLSSSFDQ--VPDELEHWVSENV-MD-AAARPWRPVSIVIEGDSRTGKTMWARSL
                  **    .  .  .  .          . ** *..** ****** *.*.*

GRD  AL1  GVHNYISGHLDFNIKSYSNNVSYNVIDDVSPTYLKLKHWKELIGAQHDWQTNCKYGKPVQ
TGMV AL1  GPHNYLSGHLDLNSRVYSNKVEYNVIDDVTPQYLKLKHWKELIGAQRDWQTNCKYGKPVQ
BGMV AL1  GPHNYLSGHLDFNSLVYSNSVEYNVIDDITPNYLKLKDWKELIGEQKDWQSNCKYGKPVQ
ACMV AL1  GPHNYLCGHLDLSPKVFNNDAWYNVIDDVDPHYLK--HFKEFMGSQRDWQSNTKYGKPVQ
ICMV AL1  GPHNYLCGHLDLSPKVYNNDAWYNVIDDVDPHYLK--HFKRIHGGPEDWQSNTKYGKPVQ
TYLCV AL1 GPHNYLCGHLDLSQKVYSNNAWYNVIDDVDPHYLK--HFKEFMGSQRDWQSNTKYGKPIQ
          * ***. ****.. ..*  ******. * ***  . *  *  ***.* *****.*

GRD  AL1  IKGGVPSILLCNPGEGSSYKDYLEGSENSALKDWTIKNAVFVFITQPMYNPQHQSS
TGMV AL1  IKGGIPSIVLCNPGEGASYKVFLDKEENTPLKNWTFHNAKFVFLNSPLYQSSTQSS
BGMV AL1  IKGGIPSIVLCNPGEGSSYKDFLNKEEKPALHNWTIHNAIFVTLTAPLYQSTAQDCQT
ACMV AL1  IKGGIPTIFLCNPGPTSSYKEFLDEEKQEALKAWALKNAIFITLTEPLYSGSNQSQSQTIQEASHPA
ICMV AL1  IKGGIPTIFLCNPGPNSSYKEFLDEEKNSALKAWALKNATFISLEGPLYSGTNQGPTQSC
TYLCV AL1 IKGGIPTIFLCNPGPQSSFKEYLDEEKNQTLKNWAIKNAIFVTIHQPLFTNTNQDPTPHRQEETSEA
          ****.*.* *****  .. *  *       *  *  . ** *     *.   . *
```

FIG. 3. Alignment of the AL1 amino acid sequences of subgroup III geminiviruses (17). BGMV, bean golden mosaic virus; ACMV, African cassave mosaic virus; ICMV, Indian casava mosaic virus; TYLCV, tomato yellow leaf and virus. Stars show identical residues; dots show conservative substitutions. Underlined sequences were pooled to create the dataset used for phylogenetic analysis. Boldface Y represents the conserved tyrosine residue present in proteins mediating initiation of rolling-circle replication (13, 14).

plasmids—but also eukaryotic geminiviral AL1 proteins (refs. 13 and 14 and references therein). All of these proteins share three similarly spaced conserved domains. In one conserved domain a tyrosine residue is present in all 59 proteins studied; this residue is at position 104 in TGMV (AL1) (15). This motif is conserved in the GRD sequences (Fig. 3). GRD also exhibits a set of a recently identified superfamily of putative NTP-binding domains found in proteins encoded by small RNA viruses and DNA viruses—e.g., SV40 T antigen—and the AL1 protein of geminiviruses. SV40 T antigen is required for SV40 replication; it has ATPase and NTP-binding activities which map to these NTP-binding motifs (15). Together these observations suggest an ancient origin for geminiviruses and, possibly, a horizontal transmission of a protogeminivirus from prokaryotes to eukaryotes.

Table 1. Similarity between tobacco and geminiviral DNA

Virus	Accession no.	% identity	% similarity	*z* score
BGMV	P05175 (Swiss)	57	70	109
TGMV	P03567 (Swiss)	57	71	112
ACMV	P14972 (Swiss)	52	69	78
ICMV	Z24758 (EMBL)	56	70	132
TYLCV	Z25751 (EMBL)	52	69	99

Geminivirus AL1 proteins retrieved from the Swiss-Prot database or deduced from the EMBL database were compared with the ancient geminivirus AL1 protein sequence. Estimates of percent identity and similarity (18) and a *z* score, to test the aligned sequences for significance of similarity, were made; $z > 10$ indicates significance (19). See legend to Fig. 3 for virus abbreviations.

It is intriguing that in GRD, sequences encoding tyrosine-104 and the putative NTP-binding site of AL1 plus the intergenic region stem and loop motifs are present. The presence of elements in GRD that are important in viral replication suggests that GRD originated from a geminiviral integration event rather than that GRDs represent protoviral sequences from which geminiviruses evolved. To determine which geminiviral lineage they are most likely to have arisen from, we used the dataset shown in Fig. 3. These pooled sequences were analyzed by methods based on the three schools of phylogenetic analysis: Parsimony, Distance Matrix, and Maximum Likelihood (21). Very similar results were obtained with all methods but only the results of Maximum Likelihood are shown (Fig. 4). The expected branches of Old World, New World, and outgroup geminiviruses were obtained, with the GRD sequences branching with New World geminiviruses, strongly suggesting that the GRD sequence is descended from an ancient New World geminivirus. It is noteworthy that *N. tabacum* is a New World plant.

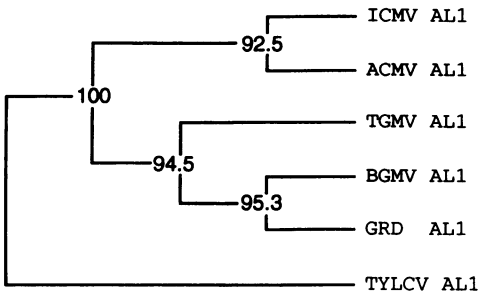

FIG. 4. Consensus of 100 Maximum Likelihood estimates of phylogeny. Here, bootstrap samples of the dataset and the DNAML program (21) were used to make 100 estimates of the phylogeny of the six sequences used. The numbers at the branch points indicate the number of times the group consisting of the species which are to the right of that fork occurred among the estimates. See legend to Fig. 3 for virus abbreviations.

DISCUSSION

The organization of GRD sequences resembles the patterns of duplications and rearrangements that occur by illegitimate recombination of artificially introduced extrachromosomal DNA within eukaryotic chromosomes. This is seen both from direct gene transfer in animal and plant systems and also, on a more limited scale, following natural gene transfer of T-DNA by the soil microorganism *Agrobacterium tumefaciens* (ref. 22 and references therein). In the case of the GRD repeats, the duplications may have occurred by local DNA replication of newly integrated viral DNA during infection. The non-GRD sequence found at the end of GRD5 may represent endogenous plant DNA which was duplicated along with GRD. Although other geminiviral sequences may also have integrated in the original recombination event, it is notable that only the cis- and trans-required replication functions appear to have been preserved. The presence of the untranscribed stem and loop also suggests recombination of geminiviral DNA with the plant genome rather than a viral transcript (*cf.* retroviruses).

As geminiviruses are not seed transmitted, the most likely mechanism for original entry into the germline is by integration into meristematic tissue which subsequently differentiated into floral tissue. The integration event may have been maintained within the species by neutral genetic drift in the small population from which tobacco evolved. Another possibility is maintenance by active selection, either by integration proximal to a favored allele, or because the integration event itself conferred a selective advantage—e.g., by alteration of expression of a neighboring host gene. Alternatively, GRD integration may have led to geminivirus resistance, by fortuitous antisense expression of the geminiviral sequence (12); or by expression of a transdominant truncated AL1 protein from GRD, or because the AL1 protein produced by superinfecting geminiviruses was effectively titrated out by the repeated GRD AL1-binding motif, either *in situ* or by transactivating release of defective interfering GRD particles (23).

Have we seen a unique event, or do geminiviruses frequently integrate into plant genomes? This could be assessed by (*i*) measuring experimentally the frequency of geminivirus integration into the plant genome and (*ii*) screening other plant species for GRD sequences. For the latter, we have performed genomic Southern blots on a dozen *Nicotiana* species, probing with GRD3-1. We find cross-hybridizing sequences only in *N. tomentosa* and *N. tomentosiformis*, two close relatives of *N. tabacum* (A. Warry, E.R.B., M.W., and C.L., unpublished work). *N. tabacum* is an amphidiploid comprising the so-called T and S genomes of the progenitors of *N. tomentosiformis* and *N. sylvestris*, respectively. It is interesting that there are no cross-hybridizing GRD sequences present in *N. sylvestris*, an observation supported by *in situ* hybridization data (16). It is difficult to determine when during evolution GRD integration took place; however, these results suggest that it may have been prior to cultivation of tobacco, as GRD sequences are also present in *N. tomentosa* and *N. tomentosiformis.*

It is interesting to view GRD repeat formation in light of another illegitimate recombination event that took place during the evolution of *Nicotiana*: that of T-DNA transformation by *Agrobacterium rhizogenes* (24, 25). It is possible that such illegitimate recombination events have been significant in the evolution of *Nicotiana.* Alternatively, such events may have been accidents that have fortuitously become fixed in the genome of tobacco because of its high developmental plasticity. This plasticity may explain why tobacco and, indeed, other *Nicotiana* species have been shown to carry horizontally acquired agrobacterial, and now geminiviral, sequences; another factor may be that tobacco has been intensively studied. Tobacco has become a favored organism for tissue culture and genetic engineering studies, in part because its developmental plasticity, which allows easy regeneration of flowering plants from meristematic somatic tissue. Perhaps other plants that propagate readily by cuttings will also be found to have incorporated exogenous DNAs. Conversely, plants that do not propagate in this way readily, such as legumes and monocots, are perhaps less likely to have acquired exogenous DNA.

Finally, we speculate, by analogy with transduction in bacteriophage infections of bacteria, and mammalian and avian oncogenic retroviruses, that gene flow may take place in both directions between plants and geminiviruses. In this way geminiviruses may be evolutionarily significant agents for horizontal gene transfer between plants. Such natural events, if evidence for them can be found, would be important in evaluating the potential consequences of the dissemination of transgenic plants and/or viruses into the environment.

We acknowledge Dr. Tony Day for the early Southern blots showing cross-hybridization of tobacco DNA with TGMV probes. We thank Dr. Douglas Maxwell and Amy Loniello for help with initial analysis of geminiviral motifs in GRD and Drs. Mike Tristem, Shirley Coomber, and Sue Cotterill for discussions and comments on the manuscript. We are also grateful to Dr. Mary-Dell Chilton for her careful editing and suggestions in revising the manuscript. This work was supported by the European Community and subsequently by the European Molecular Biology Organization (B/88000449 to E.R.B.) and a Ministry of Agriculture, Fisheries, and Food contract (RG0204) to C.L. to support M.W.

1. Berg, D. E. & Howe, M. M., eds. (1989) *Mobile DNA* (Am. Soc. Microbiol., Washington, DC).
2. Greene, A. E. & Allison, R. F. (1994) *Science* **263,** 1423–1425.
3. Xiong, Y. & Eichbush, T. H. (1990) *EMBO J.* **9,** 3353–3362.
4. Lazarowitz, S. G. (1992) *Crit. Rev. Plant Sci.* **11,** 327–349.
5. Sambrook, S., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd Ed.
6. Dellaporta, S. L., Wood, J. & Hicks, J. B. (1983) *Plant Mol. Biol. Rep.* **1,** 19–21.
7. Draper, J., Scott, R., Armitage, P. & Walden, R. (1988) *Plant Genetic Transformation and Gene Expression: A Laboratory Manual* (Blackwell, Oxford).
8. Chu, G., Vollrath, D. & Davies, R. R. W. (1986) *Science* **234,** 1582–1585.
9. Arumuganathn, K. & Earle, E. D. (1991) *Plant Mol. Biol. Rep.* **9,** 208–218.
10. Lazarowitz, S. G., Wu, L. C., Rogers, S. G. & Elmer, J. S. (1992) *Plant Cell* **4,** 799–809.
11. Fontes, E. P. B., Eagle, P. A., Sipe, P. S., Luckow, V. A. & Hanley-Bowdoin, L. (1994) *J. Biol. Chem.* **269,** 8459–8465.

12. Day, A. G., Bejarano, E. R., Burrell, M., Buck, K. & Lichtenstein, C. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 6721–6725.
13. Koonin, E. V. & Ilyina, T. V. (1992) *J. Gen. Virol.* **73,** 2763–2766.
14. Ilyina, T. V. & Koonin, E. V. (1992) *Nucleic Acids Res.* **20,** 3279–3285.
15. Gorbalenya, A. E., Koonin, E. V. & Wolf, Y. I. (1990) *FEBS Lett.* **262,** 145–148.
16. Kenton, A., Khashoggi, A., Parokonny, A., Bennett, M. D. & Lichtenstein, C. P. (1995) *Chromosome Res.* **3,** 346–350.
17. Francki, R. I. B., Fauquet, C. M., Knudson, D. L. & Brown, F. (1991) *Arch. Virol. Suppl.* **2,** 173–177.
18. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
19. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441.
20. Kornberg, A. & Baker, T. (1992) *DNA Replication* (Freeman, San Francisco), 2nd Ed.
21. Felsenstein, J. (1993) PHYLIP, Phylogeny Inference Package (Univ. Washington, Seattle), Version 3.5c.
22. Koncz, C., Németh, K., Réde, G. P. & Schell, J. (1994) in *Homologous Recombination and Gene Silencing in Plants*, ed. Paszkowski, J. (Kluwer, Dordrecht, The Netherlands), Chap. 9.
23. Stanley, J., Frischmuth, T. & Ellwood, S. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 6291–6295.
24. White, F. F., Garfinkel, D. J., Huffman, G. A., Gordon, M. P. & Nester, E. W. (1983) *Nature (London)* **301,** 348–350.
25. Furner, I. J., Huffman, G. A., Amasino, R. M., Garfinkel, D. J., Gordon, M. P. & Nester, E. W. (1986) *Nature (London)* **319,** 422–427.