A survey of geminiviruses and associated satellite DNAs in the cotton-growing areas of northwestern India

Valerio Zaffalon, Sunil Kumar Mukherjee, Vanga Siva Reddy, Jeremy R. Thompson, Mark Tepfer

Version définitive du manuscrit publié dans / Final version of the manuscript published in : Archives of Virology, 2012, 157, 483-495.
DOI: 10.1007/s00705-011-1201-y. The original publication is availalble at http://link.springer.com.

A Survey of Geminiviruses and Associated Satellite DNAs in the Cotton-Growing Areas of Northwestern India

Valerio Zaffalon[1,2], Sunil Kumar Mukherjee[3], Vanga Siva Reddy[3], Jeremy R. Thompson[1,4]* and Mark Tepfer[1,5,6]*.

[1] Plant Virology Group, ICGEB Biosafety Outstation, Via Piovega 23, 31056 Ca' Tron di Roncade, Italy.
[2] Department of Environmental Agronomy and Crop Science, University of Padova, Viale dell'Università 16, 35020 Legnaro (Padova), Italy
[3] Plant Biology Group, ICGEB Campus, Aruna Asaf Ali Marg, 110 067 New Delhi, India.
[4] Plant Science 334, Department of Plant Pathology and Plant-Microbe Biology, Cornell University, Ithaca, 14853 NY, USA.
[5] Institut Jean-Pierre Bourgin UMR1318, INRA-Versailles, 78026 Versailles Cedex, France.
[6] INRA, UR407 Pathologie Végétale, 84140 Montfavet cedex, France

*Corresponding Authors: jrt36@cornell.edu
Mark.Tepfer@versailles.inra.fr

Abstract

Severe symptoms of Cotton leaf curl disease (CLCuD) are caused by the association of a single-stranded circular DNA satellite (betasatellite) with a helper begomovirus. In this study we analyzed 40 leaf samples (primarily cotton with CLCuD symptoms and other plants growing close by) from four sites between New Delhi and the Pakistan/India border, using rolling circle amplification (RCA) and PCR. In total, the complete sequences of 12 different helper viruses, eight alphasatellites, and one betasatellite from five different plant species were obtained. A recombinant helper virusmolecule found in okra, a novel alphasatellite-related DNA from croton, and a circular DNA of 941 nucleotides isolated from clerodendrum are also described. This is the first report of the presence of both DNA components (Helper virus and betasatellite) associated with resistance-breaking CLCuD in India, and highlights the need for further work to combat its damage and spread.

Keywords: Begomovirus, Cotton leafcurl disease, Burewala, India, betasatellite, alphasatellite.

Introduction

Geminiviruses and their associated satellites present modern agriculture with important challenges, inducing a variety of diseases in tropical and sub-tropical regions [25, 30, 34]. Economic losses caused by the geminiviral cotton leaf curl disease (CLCuD) are some of the most significant of any plant virus. In the Indian subcontinent it was first identified around the Pakistani city of Multan in 1986 [16], and has since spread rapidly; in 1991 in the Pakistani Punjab the disease affected 14,000 ha out of a total growing area of 2.5 million ha, and two years later had spread to affect an overall area of 202,000 ha, resulting in a reduction in cotton yield of 29% [7]. In the following years the disease continued to spread in the Punjab and to the surrounding Pakistani regions, and in 1993 was first reported across the border in India, in Sriganganagar, Rajasthan [19]. Between 1992 and 1997, the economic loss due to CLCuD in Pakistan was calculated to be about 5 billion US dollars [7]. The establishment of the disease was assumed to be principally due to the introduction of high yielding, highly CLCuD-susceptible varieties of cotton (*Gossypium hirsutum*), in place of the native, resistant cotton species *G. arboreum*, a problem only temporarily addressed by the development of new tolerant/resistant *G. hirsutum* varieties [2-3]. Resistance subsequently

began to break down in 2001-2002, which has been ascribed to the emergence of *Cotton leaf curl Burewala virus* (CLCuBuV) and a novel recombinant betasatellite [23-24].

The etiology of CLCuD has been demonstrated to involve two infectious circular geminiviral DNAs. The first component is the helper virus (a monopartite begomovirus), which is autonomous and self-replicating, and by itself induces little or no symptoms in cotton. The second component, a betasatellite, though entirely dependent on the helper virus for replication, encapsidation, transmission by the whitefly *Bemisia tabaci*, and movement, is critical for symptom expression [8, 33]. In addition there is a third component associated with viruses causing CLCuD, an alphasatellite (also referred to as DNA 1), which depends on the helper virus for encapsidation, movement, and transmission, but in contrast to the betasatellite, replicates autonomously [22]. Alphasatellites appear to play no major role in disease maintenance or pathogenicity, although it has been demonstrated that they can reduce the accumulation the helper virus [42], possibly by competing for cellular resources which as a result attenuate the severity of the disease, playing a role similar to that of the defective interfering DNAs [9, 36]. CLCuD continues to be of primary concern to Indian and Pakistani cotton growers, and as yet there are still no positive signs that it can be effectively controlled - in Pakistan new varieties resistant (rather than tolerant) to the 2001-2002 resistance-breaking CLCuBuV have failed to materialize [2]. In India, where less work in general has been carried out on CLCuD compared to Pakistan, reports have so far identified a variety of viruses associated with the disease, namely, *Cotton leaf curl Rajasthan virus* (CLCuRaV), *Cotton leaf curl Multan virus* (CLCuMuV), *Cotton leaf curl Kokhran virus* (CLCuKoV), *Tomato leaf curl Bangalore virus* (ToLCBV), *Tomato leaf curl Bangalore virus* (ToLCBV) [1, 18-19]. This situation contrasts with the present state of the disease in Pakistan, where CLCuBuV now predominates [4].

The aim of this study was to increase our knowledge of CLCuD through a molecular epidemiology study involving the characterization of infectious circular DNA genomes present in a number of plants located in and around infected cotton fields of Northwestern India, using both PCR and rolling circle amplification (RCA). This information will improve our understanding of virus diversity and evolution, and therefore assist in the long-term in the eventual development of durable and safe resistance to the disease.

Materials and Methods

Samples

Leaves of symptomatic and asymptomatic plants were collected in Northwestern India on 17-18th September 2007, in fields in or near the cities of Hisar and Sirsa (Haryana), Kumiharwala (a village in Muktsar district, Punjab) and Sriganganagar (Rajasthan) (Fig. 1, Table 1). Incidence of CLCuD symptoms was low in cotton fields in or near Hisar, and very high at Sirsa, Kumiharwala and Sriganganagar, where it was close to 100%. In Hisar, cotton, okra and croton samples were from an agricultural context, while the other species were sampled in a garden within the city. In Sirsa, samples were from the Central Cotton Research Institute, in Kumiharwala in and adjacent to cotton fields, and in Sriganganagar from the Agriculture Research Station. Samples were stored for 1-2 months at 4°C before being shipped by courier to Ca'Tron.

DNA extraction

The DNA was extracted using a modified version of the procedure described in Doyle and Doyle (1987) [12]. 100 mg of leaf tissue per sample were used for each extraction. The CTAB extraction solution was prepared adding 2.5% PVP-40 and using 2 M NaCl and 25mM EDTA. After the addition of isopropanol, the solution was centrifuges at 10,000g for 15 min and the pellet washed with 70% ethanol. The extracted nucleic acids were respuspended in water and left at 4°C overnight. The following day, the DNA was re-extracted using one volume of cold (4°C) 1:1 phenol (pH 8.0)/chloroform, in order to get a higher purity. Its quantity and quality were measured by fluorometry. Isolate code, host, and location of all samples analyzed in this study are shown in Table 1.

Rolling circle amplification (RCA) and cloning

The GE Healthcare Illustra™ TempliPhi™ 500 Amplification Kit was used, using 20 ng of total nucleic acid extraction. The amplified samples were digested by restriction endonucleases (REs) and separated by electrophoresis on a 1.2% agarose gel in TAE. The choice of RE depended on the circular DNA of interest, with a preference for a RE recognizing a single conserved sequence in related viral sequences found in the databases. The REs of choice therefore were *Bam*HI, *Nco*I and *Sal*I for the virus and *Ava*II for betasatellite, respectively. RCA digestion products were extracted and cloned into an appropriately RE-cut plasmid (either pBluescript SKII (Stratagene, La Jolla, CA) or pGEM-T Easy (Promega, Madison, WI)) using standard cloning procedures. *Escherichia coli* (DH5α) cells were then transformed by electroporation and plasmids extracted from resultant colonies.

PCR

Two primer pairs used were designed based on Clustal alignments (Vector NTI Advance™ 9 package (Invitrogen, Carlsbad, CA)) using available database begomoviral sequences and betasatellite associated sequences. For DNA-A, PCR was carried out with degenerate primers 302 (5'-TGTGARGGYCCWTGYAARGTYCA-3') and 424 (5'-CARRTMMRRTTCAAYHACAACMTVMGGA-3') to produce an amplicon of ~827 bp. For betasatellite, PCR was carried out with primers 419 (5'-CTACCCTCCCAGGGGTACACA-3') and 420 (5'-CACGTGTTGTCATGTTGGCTT-3') to produce an amplicon of ~614 bp. Both PCR products were cleaned and sequenced.

Computer software

All sequences obtained were initially analyzed using the Vector NTI Advance™ 9 package. All phylograms were produced using PAUP* 4.0 [39] after selecting for the best nucleotide substitution method with jMODELTEST [32]. The software RDP3.27 [26] was used to analyze and detect recombination signals from DNA alignments implementing seven different methods. BLAST algorithms were used to search for sequence similarity in the GenBank (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and in the EBI databases (http://www.ebi.ac.uk/Tools/blast/). The algorithm of the NCBI Pairwise Sequence Comparison project (PASC) was used to classify the sequences taxonomically on the basis of sequence similarity [5].

Results

RCA-RFLP

RCA products were digested with *Ava*II, *Bam*HI, *Nco*I, or *Sal*I producing bands of different sizes (Full data set shown in Supplemental Table 1). In 14 cotton samples, out of a total of 23, no RCA product was obtained, which could be due either to the absence of circular DNA with the corresponding restriction site(s), or to the presence of contaminants that inhibited the formation of RCA products. Further RCA trials with diluted templates and repurified samples failed to generate restriction products (results not shown).

PCR

A PCR was considered positive if it yielded a detectable fragment of the expected size; in many cases, less intense fragments of larger and smaller sizes were also observed (results not shown). Out of a total of 40 samples tested, 36 and 35 were positive for the helper virus and betasatellite, respectively (Table 1). Only sample DhI produced no amplicons for either type of PCR. Samples 1a, 3n and HS1 were positive for betasatellite, yet negative for the helper virus; since betasatellite is dependent on a helper virus, the absence of detection of the latter is more likely due to a technical problem, rather than to the absence of the virus. Several samples (7, 10b, Cgc and RR1) were positive for the virus and negative for betasatellite, which could be absent in these samples.

Version définitive du manuscrit publié dans / Final version of the manuscript published in : Archives of Virology, 2012, 157, 483-495.
DOI: 10.1007/s00705-011-1201-y. The original publication is availalble at http://link.springer.com.

Sequencing

To confirm their viral origin, all PCR products and cloned RCA fragments were sequenced. PCR amplicons were sequenced directly: 29 from helper virus amplifications and 33 from betasatellite amplifications, of which 17 and 24 yielded readable sequences, respectively. Only a subset of sequences was readable most probably because more than one related sequence was amplified per PCR. Of the RCA-derived fragments, 15 samples yielded 44 sequences. Those shown to be full-length virus were subsequently aligned with all the geminivirus sequences in the database using the PASC algorithm [5] to determine to which viral species they belong. The main characteristics of all the sequences are shown in Table 2.

Phylogenetic analyses of the helper virus

Because of frequent observation of recombinants in begomoviruses, species attributions are made from full-length sequences, with 89% sequence identity as the species distinction cut-off (Fauquet et al. 2003). When a phylogenetic tree based on sequence similarity was generated using only the full-length sequences, all species were clearly distinguished (Fig. 2 and Supplemental Table 2). The most similar sequences in the databases were included to indicate species affiliation, and African cassava mosaic virus (ACMV, AF259894) was included as an outgroup. The relationship between *Croton yellow vein mosaic virus* (CYVMV) and *Mesta yellow vein mosaic virus* (MeYVMV) was further clarified by analysis of recombination sites, which clearly showed a tract of recombinant origin between nucleotide positions 450 and 1150 (Fig. 3). For virus and for betasatellite (see below), the sequences obtained by PCR and RCA from the same sample always grouped together, showing consistency in diagnosis (Figs 2b and 4a).

Phylogenetic analyses of satellite DNA

Overall sequence similarity among betasatellites for the partial sequences obtained was high (Supplemental Table 4). In particular, all sequences obtained from croton, papaya, guar bean, and *Tribulus terrestris* (puncture vine), as well as six sequences from tolerant and one from a susceptible cotton plant (1a), were identical and correspond to that of a CLCuM betasatellite (DQ364230). The sequences analyzed could be seen to group somewhat according to the helper viruses CLCuRaV, CLCuBuV and CYVMV (Fig. 4a), although with only partial sequences any clear pattern in grouping will not be significant. For the alphasatellites three major clades can be distinguished (Fig. 4b). If, as for betasatellites, 83% identity is used as the cutoff for distinguishing species [27], clades I and II each correspond to a species, except for FN678901 and AM711115, which have 74-75% identity with the other members of clade I. Clade II is distinct from clade I, since its members have on the order of 60% identity with clade I. One croton sample (9c-RCA-a2-F) grouped with the more unusual satellite types of clade III, whose first member was described by Saunders et al (2002) from ageratum in Singapore, and more recently described members were from Oman [17], Venezuela [35] and Brazil [31]. Clade III is only distantly related to the other two clades, with inter-clade identity levels that are equivalent to those with the non-related outgroup sequence, Coconut foliar decay virus (CFDV). Clade III is also less homogeneous, since these satellites were associated with both monopartite and bipartite begomoviruses, and the level of intra-clade sequence identity is below the 78% identity species cutoff. When the Rep protein of these satellites was compared with that of alphasatellites, nanoviruses and circoviruses, they clearly were on a separate branch from the three other groups [31, 35].

Structural features of full-length helper virus molecules

Genome structure analysis showed that all the virus molecules contain ORFs AV1, AV2, AC1, AC3 and either one of the two putative AC5 ORFs (Suppl. Fig. 1a). The ToLCNDV molecule extracted from bottle gourd sample 4 had an extra ORF, named AV3, which is not present in the other molecules. Only the CLCuBuV molecules found in cotton samples L2 and L4K lacked a complete AC2 ORF, which encodes the transcriptional activator protein, TrAP, and may also function as an

RNAi suppressor. This is also the case in the CLCuBuV reference sequence (AM421522). An alignment was performed using the three sequences found in this study and the CLCuBuV sequence (GQ247893), which was the most similar sequence present in the databases. The two single mutations at positions 1487 and 1499 (C to A and T to G, corresponding to positions 1489-1501 in CLCuBuV) created two in-frame stop codons, which shorten the ORF from 405 to 105 bp and would reduce the TrAP protein to a polypeptide of 35 aa, instead of 135.

It was suspected that molecule 10c-RCA-A1-F was the product of recombination between geminiviral DNAs, a fact confirmed with the RDP 3.27 package [26] of sequence 10c-RCA-A1-F and the reference sequences of MeYVMV (FJ159262) and *Bhendi yellow vein mosaic virus* (BYVMV) (EU589392) (Fig. 5) Two breakpoints were detected, at nucleotide positions 1108 and 1458 (encompassing the AC3 ORF (1459-1058)), with *P*-values of seven methods ranging from 4.39×10^{-04} to 1.58×10^{-11} (Supplemental Table 6). Using the BLAST algorithm a PASC value of 87.4% was obtained for 10c-RCA-A1-F with the nearest sequence, MeYVMV, therefore suggesting that in fact this virus molecule might be a putative new species.

Structural features of full-length satellite DNA sequences

Eight full-length alphasatellite-related sequences from six samples were found using RCA (Tables 1 and 2). All had the typical structure of an alphasatellite: a single ORF (coding for a replication-associated protein) and an A-rich region immediately downstream (Suppl. Fig. 1b).

Only one full-length betasatellite sequence (L2-RCA-b1-F) was obtained in this study, and was associated with CLCuBuV (Suppl. Fig. 1c). However, another circular betasatellite sequence (LG-RCA-b1-r) was only slightly shorter than L2-RCA-b1-F (1228 nts, instead of 1354 nts), and in it the region immediately upstream of the beta C1 ORF, located between nucleotide 507 and nucleotide 874, was replaced by an inverted repeat of 242 nucleotides (nts 263 to 505), which would encode most of the βC1 ORF (Suppl. Fig. 1d). The resultant molecule has an entire beta C1 ORF on the complementary sense and an almost completely duplicated beta C1 ORF on the virion sense. This is similar to another betasatellite found in the database (Acc. No. EU384591.

In order to determine whether 9c-RCA-a2 was a full-length circular DNA, PCR (not shown) with primers binding in the region flanking the *Bam*HI restriction site was performed. The PCR confirmed that the molecule was circular, but that it was 24 nucleotides longer than the cloned RCA molecule, due to the presence of an additional *Bam*HI site. Despite the low similarity to other alphasatellites, this molecule shared the same genomic structure (a single ORF in the virion sense and an A-rich region immediately downstream) and the same nona-nucleotide sequence at the putative origin of replication (TAGTATT^AC) as other alphasatellites.

Defective virus and satellite DNA molecules

RCA produced several DNA molecules of both alphasatellite (9S1-RCA-a4-D; 9S1-RCA-a5-D) and helper virus (10a1-RCA-A2-D; 10a1-RCA-A3-D; L2-RCA-A2-D; LG-RCA-A3-D) that, although the origin of replication was intact, contained a deletion in the central region. PCR using specific primers flanking the putative deleted region confirmed that the deletions were not simply artefacts (not shown). The organization of the defective virus molecules was similar, since they lacked the central region, with only certain ORFs intact, namely AV2, AV1, AC1, AC4, AC5 (Suppl. Fig. 1a). The sizes of some of the molecules were about one quarter (LG-RCA-A3-D), one third (L2-RCA-A2-D) or half (alphasatellites: 9S1-RCA-a4-D, 9S1-RCA-a5-D) of that of a full-length molecule. Molecule L2-RCA-A2-D had two deletions (confirmed also by PCR) (at positions 195-1811 and 2023-2307), and as a result, only three ORFs, a partial AV2 sequence, a partial AC1 sequence and an ORF composed of the first part of AC4 and the end of AC1.

There were three RCA-derived molecules that were helper virus recombinants, either with sequences of unknown origin (10a1-RCA-A6-r), or with viral sequences of anotherorigin (L2-RCA-A3-r, 9S1-RCA-a6-r). Molecule L2-RCA-A3-r (Suppl. Fig. 1e) was a recombinant between virus and a betasatellite, the latter being an insert in the central region. Molecule 9S1-RCA-a6-r (Suppl.

Fig. 1f), which we initially classified as an alphasatellite, is more correctly a recombinant virus sequence containing the origins of replication of both helper virus and alphasatellite.

Virus-host specificity

In general, there was a correlation between the virus species and the host species. CYVMV was detected only in croton and guar bean; the cotton leaf curl viruses were only observed in cotton; and the tomato leaf curl viruses were only in bottle gourd, cucumber, papaya and congress grass. *Clerodendrum golden mosaic China virus* is a bipartite geminivirus, the virus sequence of which was found only in the clerodendrum sample. Plants infected with more than one virus species were rare, the exceptions being okra samples 10a1 and 10c which were infected with MeYVMV and BYVMV, and croton sample 9c which was infected with CYVMV and BYVMV (and *Tomato leaf curl Patna virus* (ToLCPaV) as a partial sequence). It is of note that in one cotton sample, 9S1, three full-length alphasatellite sequences were detected, along with two defective sequences and a recombinant molecule between an alphasatellite and a CLCuV.

Discussion

The principal findings of this study are the following: 1) confirmation that CLCuBuV and associated betasatellite are present in Indian cotton fields, 2) none of the geminiviruses infecting cotton were found associated with nearby crops or weeds, 3) associated betasatellites were host-promiscuous, 4) the identification of a novel alphasatellite-related molecule. Additionally, we have demonstrated that RCA and PCR are complementary techniques (in particular when in this case RCA yields – most likely because of reduced sample quality - false negatives), and together can form a robust approach to circular DNA virus and satellite characterization.

The presence of CLCuBuV molecules signifies that this species has spread from its origin in Pakistan [4]. [24]; many of the plants from Kumiharwala and Sriganganagar, the sites closest to Pakistan, also harbor it. To our knowledge this is the first published report of the presence of this virus in India.

Sequence analysis of the CLCuBuV isolates highlights a notable truncation in the C2 gene, a homologue of the TrAP protein of bipartite begomoviruses, caused by the presence of an opal and amber stop codon. The C2 protein is multifunctional, up-regulating virion-sense gene transcription and host gene expression, while down-regulating suppression of gene silencing and the hypersensitive response [13-14, 38, 41]. More specifically, cell death elicited by the AV2 protein was shown to be suppressed by the AC2 protein [29]. The persistence of a virus with a truncated form of a gene that is clearly implicated in the modulation of plant antiviral defences is surprising, although the potential leakiness of the opal and amber stop codons [37], may point to a possible mode of gene regulation rather than complete truncation. The C-terminus functions as a transcriptional activator [10], which may or may not be functionally linked to suppression of gene silencing [41]. It may be significant that the nuclear localization signal ^{28}RRRR^{31} is just upstream of the stop codons, signifying a possible retention of gene regulatory function, although one of the residues (30) in two of isolates is a lysine instead of a arginine, a change that might affect disease phenotype [10]. Identification of a truncated TrAP gene in Pakistani CLCuV isolates has also been recently reported [4]. In that study, they showed that CLCuBuV, aside from being the only virus present in resistant cotton, possessed a C2 gene containing three mutations; facts that point possibly to more prolonged selection in favor of this virus.

One of the epidemiological aspects of CLCuD that would be useful to address is how the virus survives after cropping; it is not seed-borne and is known to have malvaceous and solanaceous hosts, including ageratum, cotton, hibiscus, okra, tobacco, and tomato. Other sources could also include weeds and cotton rattoons [7]. Although there was a limited number of cultivated crops and weeds examined in this study in none did we find viruses similar to those found in cotton, therefore suggesting a degree of host specialization, although it has been shown that the same species of begomoviruses that naturally infect okra are also found in cotton [43]. Similarly, in Burkina Faso,

Okra leaf curl diseased plants were infected by *cotton leaf curl Gezira virus* [40], while yellow vein disease of *Digera arvensis*, a common weed, was found associated with CLCuRaV [28]. Thus, while our results do not exclude the possibility transfer of CLCuV between hosts, they do not provide evidence that it is the norm.

In contrast, the picture obtained from the betasatellites points to a less selective, more promiscuous distribution. These findings are in line with previous work where seven begomoviruses associated with CLCuD have been found associated with a single betasatellite [8, 23], and Cotton leaf curl Multan betasatellite (CLCuMB) was found with yellow vein mosaic disease in mesta [11]. In our study, sequences of CLCuV-associated betasatellite were found in croton, guar bean, papaya and *Tribulus*.

All the defective molecules identified here are similar to those described previously; they contain the intergenic region and part of the AC1 gene, and are accompanied by full-length helper virus [21]. These molecules may attenuate symptoms, as has been shown for *East African cassava mosaic Cameroon virus* [6], and they could also serve as templates for recombination allowing for rapid repair of degenerate viral sequences or the creation of novel combinations of viral genes. Of the 12 full-length virus molecules sequenced in this work, only one was a novel recombinant, and its structure was the same as that of other begomoviruses described [20].

All the full-length molecules characterized here were generated by RCA, which is proving to be an invaluable tool for studying geminiviruses [15]. Our failure to obtain clonable amounts of amplified DNA from specific cotton samples by RCA, while confirming the presence of such molecules by PCR, highlights the usefulness of applying both methods for virus characterization in particular when because of less than optimal sample quality RCA yields false negatives. Cotton is a difficult material for extraction of nucleic acids, and this problem might have limited the efficiency of RCA in our lab. With RCA-negative samples, even when the extracted nucleic acids were diluted to avoid the possible harmful effects of contaminating substances [15], the RCA experiments remained unsuccessful. Our results clearly show that, at least for detection purposes, a good set of PCR primers is sufficient although in most cases the partial sequence produced cannot be used to conclusively identify the DNA component.

The results of this study have unequivocally demonstrated the presence of the resistance-breaking CLCuBuV in India, a fact that can be in part explained by the virus' initial rapid spread from its source in Pakistan, assisted by wind conditions that would favor the movement of *Bemisia tabaci*, and the nature of geography of the border region. Recent work in breeding has so far not provided a complete solution to this threat to cotton cultivation [2], therefore highlighting the urgent need for a more concerted effort to understand the ecology of CLCuD-associated viruses and their control.

Acknowledgements

Many thanks to Dr. Dilip Mongia, CICR Sirsa, Prof. Narayan Rishi, CCS Agriculture University, Hissar for assistance in the field, and to Mirco Boem, Laura Chiappetta, Slavica Matic, Marco Morroni, Arianna Friscina and Amit Bhardwaj for technical support and sound advice. This work was funded in part by Fondazione Cassamarca. Financial support of ICGEB is acknowledged.

References

1. Ahuja SL, Monga D, Dhayal LS (2007) Genetics of resistance to cotton leaf curl disease in Gossypium hirsutum L. under field conditions. J Hered 98:79-83
2. Akhtar KP, Jamil FF, Haq MA, Khan IA (2008) Comparison of resistance to cotton leaf curl disease (Multan/Burewala) in *Gossypium hirsutum* L. varieties and breeding lines. J Phytopathology 156:352–357
3. Ali M (1997) Breeding of cotton varieties for resistance to cotton leaf curl virus. Pakistan J Phytopathol 9:1-7

4. Amrao L, Amin I, Shahid MS, Briddon RW, Mansoor S (2010) Cotton leaf curl disease in resistant cotton is associated with a single begomovirus that lacks an intact transcriptional activator protein. Virus Res
5. Bao Y, Kapustin Y, Tatusova T (2008) Virus Classification by Pairwise Sequence Comparison (PASC). . In: Mahy BWJ, Van Regenmortel MHV (eds) Encyclopedia of Virology. Elsevier, Oxford, pp 342-348.
6. Behjatnia SA, Dry IB, Rezaian MA (2007) Characterization and transient replication of tomato leaf curl virus defective DNAs. Arch Virol 152:1127-1138
7. Briddon RW, Markham PG (2000) Cotton leaf curl virus disease. Virus Res 71:151-159
8. Briddon RW, Mansoor S, Bedford ID, Pinner MS, Saunders K, Stanley J, Zafar Y, Malik KA, Markham PG (2001) Identification of dna components required for induction of cotton leaf curl disease. Virology 285:234-243
9. Briddon RW, Stanley J (2006) Subviral agents associated with plant single-stranded DNA viruses. Virology 344:198-210
10. Chowda-Reddy RV, Dong W, Felton C, Ryman D, Ballard K, Fondong VN (2009) Characterization of the cassava geminivirus transcription activation protein putative nuclear localization signal. Virus Res 145:270-278
11. Das S, Roy A, Ghosh R, Paul S, Acharyya S, Ghosh SK (2008) Sequence variability and phylogenetic relationship of betasatellite isolates associated with yellow vein mosaic disease of mesta in India. Virus Genes 37:414-424
12. Doyle JJ, Doyle JL (1987) A rapid DNA isolation method for small quantities of fresh tissues. Phytochemical Bulletin of the Botanical Society of America 19:11-15
13. Dry I, Krake L, Mullineaux P, Rezaian A (2000) Regulation of tomato leaf curl viral gene expression in host tissues. Mol Plant Microbe Interact 13:529-537
14. Gopal P, Pravin Kumar P, Sinilal B, Jose J, Kasin Yadunandam A, Usha R (2007) Differential roles of C4 and betaC1 in mediating suppression of post-transcriptional gene silencing: evidence for transactivation by the C2 of Bhendi yellow vein mosaic virus, a monopartite begomovirus. Virus Res 123:9-18
15. Haible D, Kober S, Jeske H (2006) Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J Virol Methods 135:9-16
16. Hussain T, Mahmood T (1988) A note on leaf curl disease of cotton. Pakistan Cottons 32(4):248-251
17. Idris AM, Shahid MS, Briddon RW, Khan AJ, Zhu JK, Brown JK (2011) An unusual alphasatellite associated with monopartite begomoviruses attenuates symptoms and reduces betasatellite accumulation. J Gen Virol 92:706-717
18. Kirthi N, Priyadarshini CG, Sharma P, Maiya SP, Hemalatha V, Sivaraman P, Dhawan P, Rishi N, Savithri HS (2004) Genetic variability of begomoviruses associated with cotton leaf curl disease originating from India. Arch Virol 149:2047-2057
19. Kumar A, Kumar J, Khan JA (2010) Sequence characterization of cotton leaf curl virus from Rajasthan: phylogenetic relationship with other members of geminiviruses and detection of recombination. Virus Genes 40:282-289
20. Lefeuvre P, Martin DP, Hoareau M, Naze F, Delatte H, Thierry M, Varsani A, Becker N, Reynaud B, Lett JM (2007) Begomovirus 'melting pot' in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. J Gen Virol 88:3458-3468
21. Liu Y, Robinson DJ, Harrison BD (1998) Defective forms of cotton leaf curl virus DNA-A that have different combinations of sequence deletion, duplication, inversion and rearrangement. J Gen Virol 79 (Pt 6):1501-1508
22. Mansoor S, Khan SH, Bashir A, Saeed M, Zafar Y, Malik KA, Briddon R, Stanley J, Markham PG (1999) Identification of a novel circular single-stranded DNA associated with cotton leaf curl disease in Pakistan. Virology 259:190-199

23. Mansoor S, Briddon RW, Bull SE, Bedford ID, Bashir A, Hussain M, Saeed M, Zafar Y, Malik KA, Fauquet C, Markham PG (2003) Cotton leaf curl disease is associated with multiple monopartite begomoviruses supported by single DNA beta. Arch Virol 148:1969-1986
24. Mansoor S, Amin I, Hussain M, Zafar Y, Briddon RW (2006) Engineering novel traits in plants through RNA interference. Trends Plant Sci 11:559-565
25. Mansoor S, Zafar Y, Briddon RW (2006) Geminivirus disease complexes: the threat is spreading. Trends Plant Sci 11:209-212
26. Martin DP, Williamson C, Posada D (2005) RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21:260-262
27. Mubin M, Briddon RW, Mansoor S (2009) Complete nucleotide sequence of chili leaf curl virus and its associated satellites naturally infecting potato in Pakistan. Arch Virol 154:365-368
28. Mubin M, Briddon RW, Mansoor S (2009) Diverse and recombinant DNA betasatellites are associated with a begomovirus disease complex of Digera arvensis, a weed host. Virus Res 142:208-212
29. Mubin M, Amin I, Amrao L, Briddon RW, Mansoor S (2010) The hypersensitive response induced by the V2 protein of a monopartite begomovirus is countered by the C2 protein. Mol Plant Pathol 11:245-254
30. Nawaz-ul-Rehman MS, Fauquet CM (2009) Evolution of geminiviruses and their satellites. FEBS Lett 583:1825-1832
31. Paprotka T, Metzler V, Jeske H (2010) The first DNA 1-like alpha satellites in association with New World begomoviruses in natural infections. Virology 404:148-157
32. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253-1256
33. Qazi J, Amin I, Mansoor S, Iqbal MJ, Briddon RW (2007) Contribution of the satellite encoded gene betaC1 to cotton leaf curl disease symptoms. Virus Res 128:135-139
34. Rojas MR, Hagen C, Lucas WJ, Gilbertson RL (2005) Exploiting chinks in the plant's armor: evolution and emergence of geminiviruses. Annu Rev Phytopathol 43:361-394
35. Romay G, Chirinos D, Geraud-Pouey F, Desbiez C (2010) Association of an atypical alphasatellite with a bipartite New World begomovirus. Arch Virol 155:1843-1847
36. Saunders K, Bedford ID, Briddon RW, Markham PG, Wong SM, Stanley J (2000) A unique virus complex causes Ageratum yellow vein disease. Proc Natl Acad Sci U S A 97:6890-6895
37. Skuzeski JM, Nichols LM, Gesteland RF, Atkins JF (1991) The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. J Mol Biol 218:365-373
38. Sunter G, Bisaro DM (1992) Transactivation of geminivirus AR1 and BR1 gene expression by the viral AL2 gene product occurs at the level of transcription. Plant Cell 4:1321-1331
39. Swofford DL (1993) Paup - a Computer-Program for Phylogenetic Inference Using Maximum Parsimony. J Gen Physiol 102:A9-A9
40. Tiendrebeogo F, Lefeuvre P, Hoareau M, Villemot J, Konate G, Traore AS, Barro N, Traore VS, Reynaud B, Traore O, Lett JM (2010) Molecular diversity of cotton leaf curl Gezira virus isolates and their satellite DNAs associated with okra leaf curl disease in Burkina Faso. Virol J 7:48
41. Trinks D, Rajeswaran R, Shivaprasad PV, Akbergenov R, Oakeley EJ, Veluthambi K, Hohn T, Pooggin MM (2005) Suppression of RNA silencing by a geminivirus nuclear protein, AC2, correlates with transactivation of host genes. J Virol 79:2517-2527
42. Wu P-J, Zhou X-P (2005) Interaction between a Nanovirus-like Component and the Tobacco Curly Shoot Virus/Satellite Complex. Acta Biochimica et Biophysica Sinica 37:25-31
43. Zhou X, Liu Y, Robinson DJ, Harrison BD (1998) Four DNA-A variants among Pakistani isolates of cotton leaf curl virus and their affinities to DNA-A of geminivirus isolates from okra. J Gen Virol 79 (Pt 4):915-923

Version définitive du manuscrit publié dans / Final version of the manuscript published in : Archives of Virology, 2012, 157, 483-495.
DOI: 10.1007/s00705-011-1201-y. The original publication is availalble at http://link.springer.com.

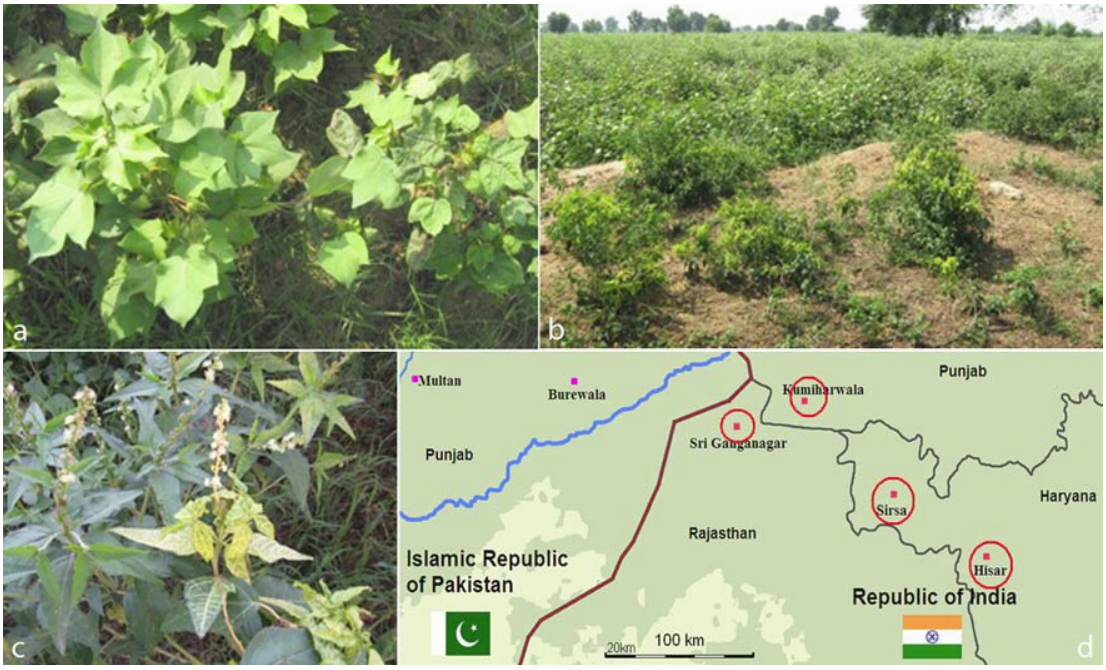


Fig. 1. Plant symptoms and sampling locations. Cotton and croton plants in a field in Northwestern India. (a) a healthy cotton plant (left) and a cotton plant affected by leaf curl disease (right). (b) a typical Northwestern Indian cotton field environment with infected croton (foreground) growing next to a cotton field. (c) a healthy croton plant (left) and a croton plant affected by yellow vein mosaic disease (right). (d) map of the sampling region. The cities where the samples were collected are shown along with the cities of Multan (Pakistan), where CLCuD was originally detected, and Burewala from which currently the most dominant species takes its name.

Version définitive du manuscrit publié dans / Final version of the manuscript published in : Archives of Virology, 2012, 157, 483-495.
DOI: 10.1007/s00705-011-1201-y. The original publication is availalble at http://link.springer.com.

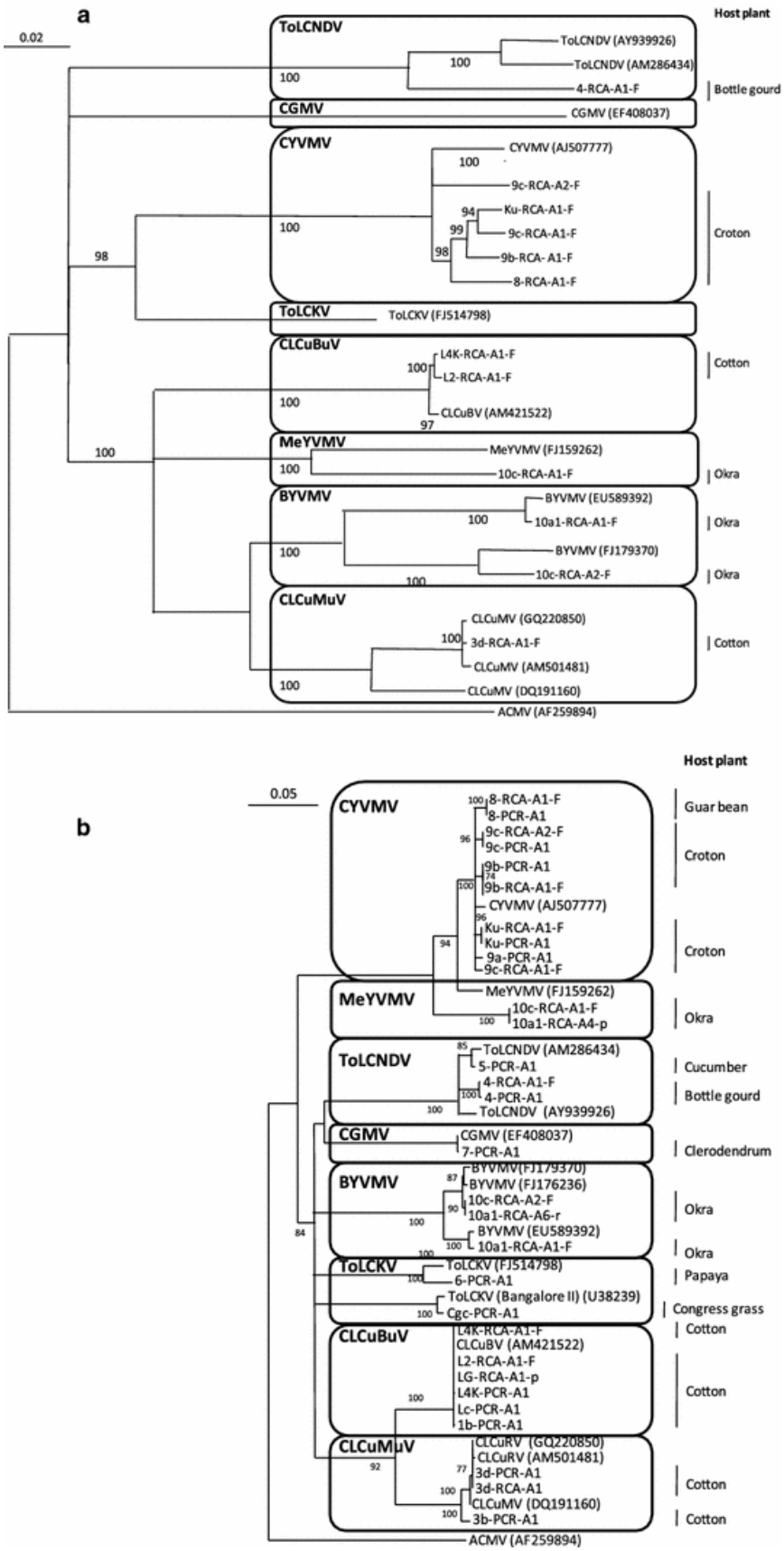


Fig. 2. Phylogenetic trees of helper virus sequences (a) full-length sequences and (b) 492-498 nts in the AV1 ORF. Trees were obtained obtained using PAUP*4.0 (neighbor-joining) (Swofford, 1993) after selecting for the best nucleotide substitution model with jModeltest (Posada, 2008). Bootstrap values were calculated from 100 iterations. Branches with bootstrap values less than 70 were collapsed. Bar: number of substitutions per nucleotide. Virus species in bold. ACMV was used as outgroup. TLCuNDV – *Tomato leaf curl New Delhi virus*, CGMV - *Clerodendrum golden mosaic China virus*, ToLCKV - *Tomato leaf curl Karnataka virus*

Version définitive du manuscrit publié dans / Final version of the manuscript published in : Archives of Virology, 2012, 157, 483-495.
DOI: 10.1007/s00705-011-1201-y. The original publication is availalble at http://link.springer.com.

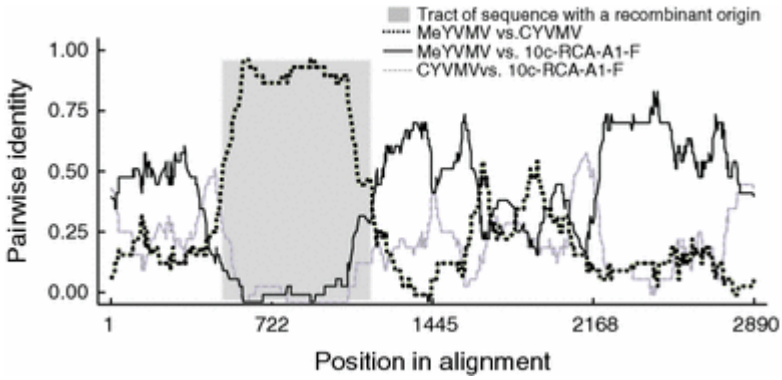


Fig. 3. Recombination involving MeYVMV and CYVMV. Pairwise identities (y-axis) between MeYVMV, CYVMV and molecule 10c-RCA-A1-F as determined by the RDP3 program (Martin et al., 2005). Nucleotide position in the genome shown on x-axis.

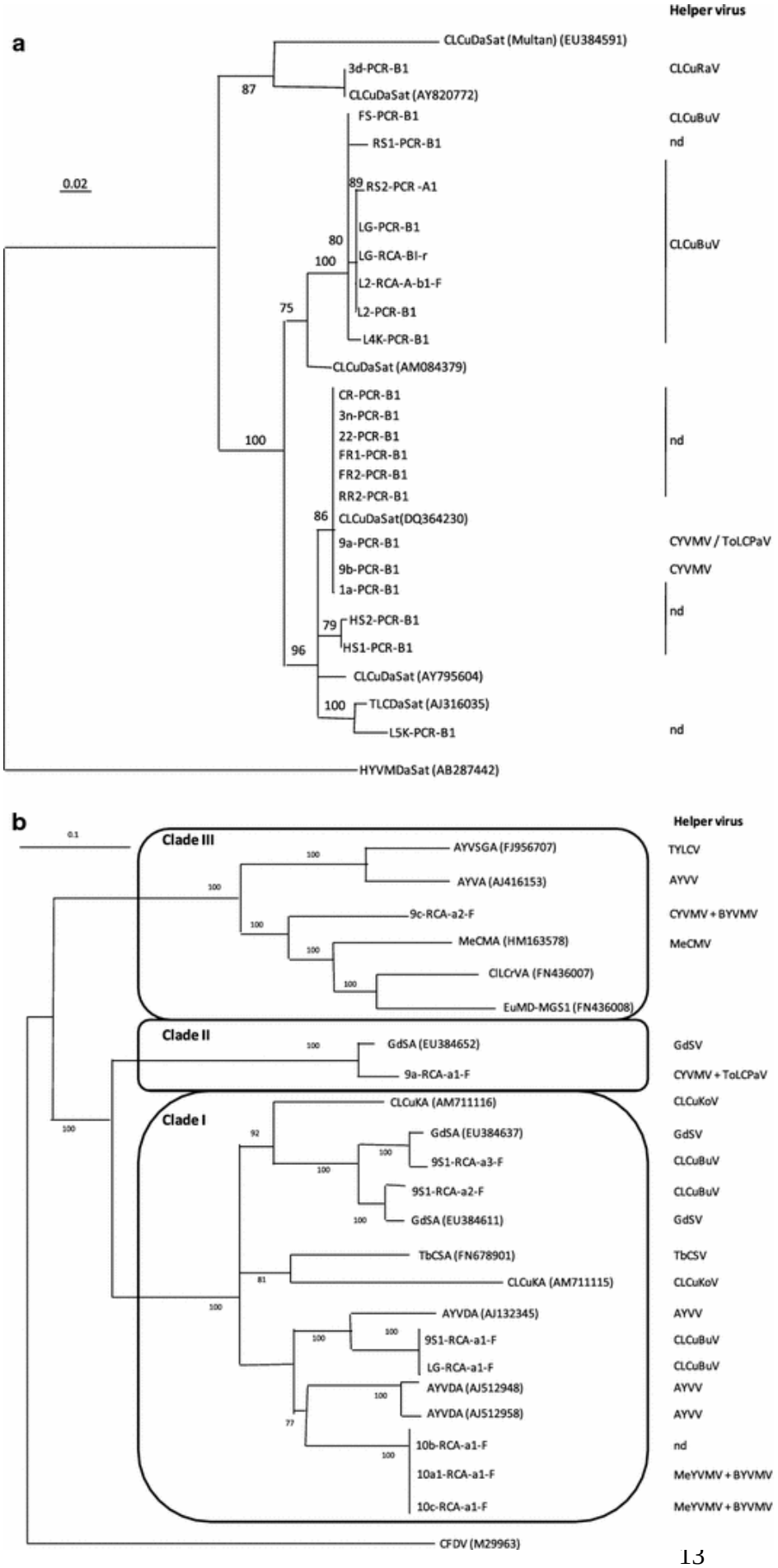


Fig. 4. Phylogenetic trees of satellite DNA sequences. (a) betasatellite sequences (274-293 nts including the A-rich region), and (b) full-length alphasatellite sequences. The trees were obtained using PAUP*4.0 (neighbor-joining) [39] after selecting for the best nucleotide substitution model with jModeltest [32]. Bootstrap values were calculated from 100 iterations. Branches with bootstrap values less than 70 were collapsed. Bar: number of substitutions per nucleotide. HYVMVA and CFDV were used as outgroups in (a) and (b), respectively. Nd- not defined. AYVV- *Ageratum yellow vein virus*, MeCMV - *melon chlorotic mosaic virus*, GdSA - Gossypium davidsonii symptomless alphasatellite, TbCSV - *Tobacco curly shoot virus.*

Version définitive du manuscrit publié dans / Final version of the manuscript published in : Archives of Virology, 2012, 157, 483-495.
DOI: 10.1007/s00705-011-1201-y. The original publication is availalble at http://link.springer.com.

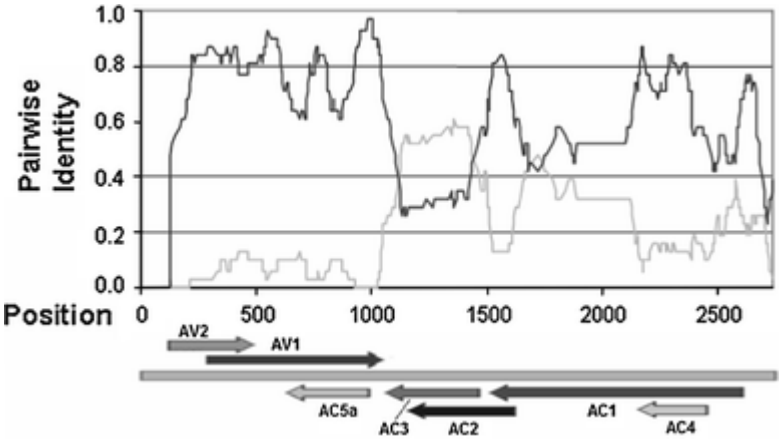


Fig. 5. Pairwise identities (y-axis) between molecule 10c-RCA-A1-F and EU589392 (light grey) and between 10c-RCA-A1-F and FJ159262 (dark grey) in relation to nucleotide position (x-axis) in the genome as determined by the RDP3 program [26]. Below the graph, the genome organization of 10c-RCA-A1-F is displayed with the direction of each ORF indicated.

Version définitive du manuscrit publié dans / Final version of the manuscript published in : Archives of Virology, 2012, 157, 483-495. DOI: 10.1007/s00705-011-1201-y. The original publication is availalble at http://link.springer.com.

Table 1. Symptoms and presence/absence of geminiviral DNAs in samples analysed

Sample	Location	Host species	Symptoms	PCR		RCA
				A	*β*	
9a	Hisar	Croton (*Croton sp.*)	+	+	+	A, α
9b	Hisar	Croton	+	+	+	A
9c	Hisar	Croton	+	+	+/-	A, α
4	Hisar	Bottle gourd (*Lagenaria siceraria*)	+	+	+/-	A
7	Hisar	Clerodendrum (*Clerodendrum sp.*)	+	+	-	A, d
1a	Hisar	Cotton (*Gossypium hirsutum*)	+	-	+	n.a.
1b	Hisar	Cotton	+	+	+	+
3b	Hisar	Cotton	+	+	+	n.a.
3d	Hisar	Cotton	+	+	+	A
3n	Hisar	Cotton	-	-	+	+
5	Hisar	Cucumber (*Cucumis sativus*)	+	+	+	+
8	Hisar	Guar bean (*Cyamopsis tetragonoloba*)	+	+	+	A
10a	Hisar	Okra (*Abelmoschus esculentus*)	+	+	+	A, α
10b	Hisar	Okra	+	+/-	-	α
10c	Hisar	Okra	+	+	+	A, α
6	Hisar	Papaya (*Carica papaya*)	+	+	+	+
HS1	Sirsa	Cotton	+	-	+/-	+
HS2	Sirsa	Cotton	+	+/-	+	n.a.
Ku	Kumiharwala	Croton	+	+	+	A
Cg	Kumiharwala	Congress grass (*Parthenium hysterophorus*)	+	+/-	+	+
Cgc	Kumiharwala	Congress grass	-	+/-	-	+
CR	Kumiharwala	Cotton	-	+/-	+/-	n.a.
L2	Kumiharwala	Cotton	+	+	+	A, β
L4K	Kumiharwala	Cotton	+	+	+	A
L5K	Kumiharwala	Cotton	+	+	+	n.a.
Lc	Kumiharwala	Cotton	-	+	+	n.a.
LG	Kumiharwala	Cotton	+	+	+	A, α, β
130 (130/1)	Sriganganagar	Cotton	-	+	+	n.a.
22 (22/2)	Sriganganagar	Cotton	-	+	+	+
FR1 (F/861)	Sriganganagar	Cotton	-	+	+	n.a.
FR2 (F/861)	Sriganganagar	Cotton	-	+	+	n.a.
FS	Sriganganagar	Cotton	+	+	+	n.a.
RR2 (RS2013)	Sriganganagar	Cotton	-	+/-	+	n.a.
RR1 (RS2013)	Sriganganagar	Cotton	-	+/-	-	n.a.
RS1	Sriganganagar	Cotton	+	+	+	n.a.
RS2	Sriganganagar	Cotton	+	+	+	n.a.
9S1	Sriganganagar	Cotton	+	+	+	α
DhI	Sriganganagar	Digera (*Digera sp.*)	+	-	-	+
Eup	Sriganganagar	Euphorbia (*Euphorbia sp.*)	-	+	+	n.a.
Tri	Sriganganagar	Puncture vine *(Tribulus terrestris)*	-	+/-	+	+

From left to right the columns indicate: name, location, host plant, + or – symptoms, + or - betasatellite in dot-blot, DNAs detected by PCR (A, DNA-A; β, betasatellite. +/- indicates samples that were positive in some but not all tests); DNAs detected by RCA, (A, DNA-A; α, alphasatellite; β, betasatellite; d, other molecule; +, RCA product observed, sequence not obtained; n.a., sample not amplified). Highlighted PCR or RCA products were sequenced.

Table 2. Characteristics of viral sequences detected by PCR and RCA.

Sample	Host plant	Sequence name	Accession number	Length	Species assignment	Closest hit	Accession number of closest hit	Coverage (%)	Max. identity (%)	PASC (%)
9a	Croton	9a-PCR-A1	FN645895	586	n.a.	CYVMV	AJ507777.1	100	99	
		9a-RCA-A1-p	FN645896	1814	n.a.	ToLCPaV	EU862323.1	100	87	
		9a-RCA-a1-F	FN658709	1372	n.a.	GdSA	EU384652.1	88	91	
		9a-PCR-B1	-	499	n.a.	CLCuMuB	DQ364230.1	100	100	
9b	Croton	9b-PCR-A1	FN645897	676	n.a.	CYVMV	AJ507777.1	99	99	
		9b-RCA-A1-F	FN645898	2760	CYVMV	n.a.	AJ507777.1			95.6
		9b-PCR-B1	-	546	n.a.	CLCuMuB	DQ364230.1	100	99	
9c	Croton	9c-PCR-A1	FN645900	742	n.a.	CYVMV	AJ507777.1	100	99	
		9c-RCA-A1-F	FN645901	2749	CYVMV	n.a.	AJ507777.1	100	95	
		9c-RCA-A2-F	FN645902	2750	CYVMV	n.a.	AJ507777.1	98	95	
		9c-RCA-A3-p	FN645903	1558	n.a.	BYVMV	AJ002453.1	100	96	
		9c-RCA-a1-p	FN658710	1214	n.a.	Alphasatellite	AJ512958.1	90	97	
		9c-RCA-a2-F	FN658711	1347	n.a.	AYVV alphasatellite	AJ416153.1	31	67	
4	Bottle gourd	4-PCR-A1	FN645904	675	n.a.	ToLCNDV	AY939926.1	100	96	
		4-RCA-A1-F	FN645905	2737	ToLCNDV	n.a.	AY939926.1			89.8
		4-RCA-A2-p	FN645906	701	n.a.	ToLCNDV	DQ629102.1	100	96	
7	Clerodendrum	7-PCR-A1	FN645907	739	n.a.	ClGMV	EF408037.1	100	99	
		7-RCA-A1-p	FN645908	527	n.a.	ClGMV	EF408037.1	100	99	
		7-RCA-d1-F	FN658712	941	n.a.	No similarity found				
1a	Cotton	1a-PCR-B1	-	394	n.a.	CLCuMuB	DQ364230.1	100	100	
1b	Cotton	1b-PCR-A1	FN645909	663	n.a.	CLCuBuV	AM421522.1	100	99	
		1b-PCR-B1	FN658713	284	n.a.	CLCuMuB	AY795608.2	100	98	
3n	Cotton	3n-PCR-B1	-	535	n.a.	CLCuMuB	DQ364230.1	100	100	
3b	Cotton	3b-PCR-A1	FN645910	521	n.a.	CLCuMuV	DQ191160.1	100	98	
		3b-PCR-B1	FN658714	232	n.a.	CLCuMuB	AY795608.2	100	98	
3d	Cotton	3d-PCR-A1	FN645911	664	n.a.	CLCuRaV	GQ220850.1	100	100	
		3d-RCA-A1-F	FN645912	2753	CLCuRaV	n.a.	AM501481.1			99.6
		3d-PCR-B1	FN658715	269	n.a.	CLCuMuB	AY820772.1	100	100	
5	Cucumber	5-PCR-A1	FN645913	655	n.a.	ToLCNDV	AM286434.1	100	98	
8	Guar bean	8-PCR-A1	FN645914	639	n.a.	CYVMV	AJ507777.1	99	97	
		8-RCA-A1-F	FN645915	2755	CYVMV	n.a.	AJ507777.1			94.7
		8-RCA-A2-p	FN645916	754 + 650	n.a.	ToLCNDV	AY428769.1	100	98	
		8-PCR-B1	-	400	n.a.	CLCuMuB	DQ364230.1	100	100	

Sample	Host plant	Sequence name	Accession number	Length	Species assignment	Closest hit	Accession number of closest hit	Coverage (%)	Max. identity (%)	PASC (%)
10a1	Okra	10a1-RCA-A1-F	FN645917	2744	BYVMV	n.a.	EU589392.1			99.0
		10a1-RCA-A2-D	FN645918	2171	n.a.	MeYVMV	FJ159262.1	96	89	
		10a1-RCA-A3-D	FN645919	1102	n.a.	MeYVMV	FJ159262.1	94	87	
		10a1-RCA-A4-p	FN645920	1638	n.a.	MeYVMV	FJ159262.1	100	91	
		10a1-RCA-A5-p	FN645921	1109	n.a.	BYVMV	AJ002453.1	99	91	
		10a1-RCA-A6-r	-	2666	n.a.	BYVMV	FJ179370.1	88	100	
		10a1-RCA-a1-F	FN658716	1377	n.a.	Alphasatellite	AJ512948.1	100	85	
10b	Okra	10b-RCA-a1-F	FN658718	1377	n.a.	Alphasatellite	AJ512948.1	100	85	
10c	Okra	10c-RCA-A1-F	FN645922	2742	'MeYVMV'	n.a.	FJ159262.1			87.4
		10c-RCA-A2-F	FN645923	2709	n.a.	BYVMV	FJ176236.1	95	96	100
		10c-RCA-a1-F	FN658717	1377	n.a.	Alphasatellite	AJ512948.1	100	85	
6	Papaya	6-PCR-A1	FN645924	749	n.a.	ToLCKaV	FJ514798.1	100	94	
		6-PCR-B1	-	542	n.a.	CLCuMuB	DQ364230.1	100	100	
HS1	Cotton	HS1-PCR-B1	FN658720	515	n.a.	CLCuMuB	AY795604.1	100	100	
HS2	Cotton	HS2-PCR-B1	FN658719	569	n.a.	CLCuMuB	AY795604.1	100	100	
Ku	Croton	Ku-PCR-A1	FN645925	558	n.a.	CYVMV	AJ507777.1	100	98	
		Ku-RCA-A1-F	FN645926	2761	CYVMV	n.a.	AJ507777.1			95.8
		Ku-RCA-A2-p	FN645927	629	n.a.	CYVMV	AJ507777.1	100	94	
Cgc	Congress grass	Cgc-PCR-A1	FN645928	659	n.a.	ToLCKaV	U38239.1	100	98	
CR	Cotton	CR-PCR-B1	-	516	n.a.	CLCuMuB	DQ364230.1	100	100	
L2	Cotton	L2-RCA-A1-F	FN645929	2759	CLCuBuV	n.a.	AM421522.1			99.1
		L2-RCA-A2-D	FN645930	859	n.a.	CLCuBuV	AM421522.1	100	100	
		L2-RCA-A3-r	FN658736	827	n.a.					
				(642)		CLCuBuV	AM421522.1	100	99	
				(192)		CLCuMuB	AY795604.1	99	98	
		L2-PCR-B1	FN658721	526	n.a.	CLCuMuB	AM084379.1	100	96	
		L2-RCA-b1-F	FN658722	1354	n.a.	CLCuMuB	AM084379.1	100	96	
L4K	Cotton	L4K-PCR-A1	FN645931	609	n.a.	CLCuBuV	AM421522.1	100	100	
		L4K-RCA-A1-F	FN645932	2759	CLCuBuV	n.a.	AM421522.1			99.2
		L4K-PCR-B1	FN658723	568	n.a.	CLCuMuB	AM084379.1	100	100	
L5K	Cotton	L5K-PCR-B1	FN658724	566	n.a.	CLCuMuB	AJ316035.1	100	100	
Lc	Cotton	Lc-PCR-A1	FN645933	740	n.a.	CLCuBuV	AM421522.1	100	100	
		Lc-PCR-B1	FN658725	205	n.a.	CLCuMuB	AY795604.1	100	99	
LG	Cotton	LG-RCA-A1-p	FN645934	1182	n.a.	CLCuBuV	AM421522.1	100	99	
		LG-RCA-A2-p	FN645935	1583	n.a.	CLCuBuV	AM421522.1	100	99	

Sample	Host plant	Sequence name	Accession number	Length	Species assignment	Closest hit	Accession number of closest hit	Coverage (%)	Max. identity (%)	PASC (%)
		LG-RCA-A3-D	FN645936	646	n.a.	CLCuBuV	AM421522.1	100	99	
		LG-RCA-a1-F	FN658727	1366	n.a.	Alphasatellite	AJ132345.1	100	99	
		LG-RCA-b1-r	FN658737	1228	n.a.	CLCuMuB	EU384591.1	95	97	
		LG-PCR-B1	FN658726	562	n.a.	CLCuMuB	AM084379.1	100	96	
22	Cotton	22-PCR-B1	-	594	n.a.	CLCuMuB	DQ364230.1	99	100	
FR1	Cotton	FR1-PCR-B1	-	465	n.a.	CLCuMuB	DQ364230.1	99	100	
PR2	Cotton	PR2-PCR-B1	-	428	n.a.	CLCuMuB	DQ364230.1	100	99	
RR2	Cotton	RR2-PCR-B1	-	611	n.a.	CLCuMuB	DQ364230.1	99	100	
RS1	Cotton	RS1-PCR-B1	FN658732	531	n.a.	CLCuMuB	AM084379.1	100	100	
RS2	Cotton	RS2-PCR-A1	FN645937	616	n.a.	CLCuBuV	AM421522.1	100	100	
		RS2-PCR-B1	FN658731	527	n.a.	CLCuMuB	AM084379.1	100	96	
9S1	Cotton	9S1-PCR-A1	FN645938	349	n.a.	CLCuBuV	AM421522.1	100	94	
		9S1-RCA-a1-F	FN658728	1365	n.a.	Alphasatellite	AJ132345.1	100	89	
		9S1-RCA-a2-F	FN658729	1369	n.a.	GdSA	EU384611.1	100	97	
		9S1-RCA-a3-F	FN658730	1378	n.a.	GdSA	EU384637.1	100	96	
		9S1-RCA-a4-D	FN658733	694	n.a.	GdSA	EU384637.1	100	98	
		9S1-RCA-a5-D	FN658734	636	n.a.	GdSA	EU384637.1	100	98	
		9S1-RCA-a6-r	FN658735	1287	n.a.	GdSA				
				(875)			EU384637.1	100	98	
				(419)		CLCuBuV	AJ228582.1	100	100	
Tri	*Tribulus*	Tri-PCR-B1	-	532	n.a.	CLCuMuB	DQ364230.1	100	100	

The sequence names indicate: sample, technique used (PCR or RCA), type of molecule (A: DNA-A, B: betasatellite, a: alphasatellite, b: betasatellite), identification number, sequence description (F: full-length, D: defective, r: DNA-A recombinant, p: partially sequenced. The length of the sequence is displayed along with the most similar sequence found using blastn tool against GenBank database sequences: the accession number of the best hit, along with the coverage and the identities of the alignment and the score of PASC alignment. A dash in the accession number column means that the sequence was identical to another submitted sequence. GdSA - Gossypium davidsonii symptomless alphasatellite, ClGMV - *Clerodendron yellow mosaic virus*

Supplemental Table 1. Fragment lengths of restricted RCA products. RCA products of all samples were digested with restriction enzymes. Fragment lengths were calculated by the Biorad Quantity One ® software and rounded to the nearest 50 bases. Fragments in parentheses were cloned together. A more concise version of this table showing only those fragment lengths cloned and sequenced is shown in table 2.

	Host species	*Ava*II	*Bam*HI	*Nco*I	*Sal*I
9a	Croton	1750 1550 1300 1050 900 400 200	2700 1200 700 600	N.P.	2800
9b	Croton	1450 800 650 350 250	2600 1150 700 650 250	N.P.	2750
9c	Croton	1550 900 850 600 450 250	2750 1300 1200 800	N.P.	2700
4	Bottle Gourd	1300 900 850 750 600 250	2750 2100	N.P.	2700
7	Clerodendrum	1050 800 500 450	2100 (1350 1100) 850 300	N.P.	900
1a	Cotton	N.A.	N.A.	N.P.	N.P.
1b	Cotton	1300 900	1650 1250	N.P.	N.P.
3n	Cotton		2850	N.P.	N.P.
3d	Cotton	1350 1150 900 550 400	2750 1550 1350 1150 700	N.P.	2800 1400
3b	Cotton	N.A.	N.A.	N.P.	N.P.
5	Cucumber	1200 900 650	N.A.	N.P.	N.P.
8	Guar Bean	1550 1190 900	2750 2100 1350 700	N.P.	2650
10a	Okra	1250 1000 900 750 600 450	2650 (1550 1350) 1200	N.P.	1600 1050
10b	Okra	1500 1200 950 850 800 650 600 350 150 50	2700 (1650 1450) 1300	N.P.	1700 1150
10c	Okra	1250 1050 900	2700 (1550 1350) 1150	N.P.	1650 1050
6	Papaya	1050 950 800 700 650 550 500 200	N.A.	N.P.	N.P.
HS2	Cotton	N.A.	N.A.	N.A.	N.A.
HS1	Cotton	1000 600	N.A.	N.P.	N.P.
Ku	Croton	1550 1300 900 800 400	2750	N.P.	2650
Cg	Congress grass	1450	N.A.	N.P.	N.P.
Cgc	Congress grass	1250 1100	2900	N.P.	N.P.
CR	Cotton	N.A.	N.A.	N.P.	N.P.
L2	Cotton	1350 1050 850 550 350 300	1600 1200 900 650	2550	2500 1250
L4K	Cotton	1400 1050 850 550 200	1650 1250	2450	2600 1250
L5K	Cotton	N.A.	N.A.	N.A.	N.A.
Lc	Cotton	N.A.	N.A.	N.P.	N.P.
LG	Cotton	1250 1000 900 750 600 500	1600 (1350 1200) 650	N.P.	2900 1400
130	Cotton	N.A.	N.A.	N.P.	N.P.
22	Cotton	N.A.	2800	N.P.	N.P.
FR1	Cotton	N.A.	N.A.	N.P.	N.P.

FR2	Cotton	N.A.	N.A.	N.P.	N.P.
FS	Cotton	N.A.	N.A.	N.P.	N.P.
RR1	Cotton	N.A.	N.A.	N.P.	N.P.
RR2	Cotton	N.A.	N.A.	N.P.	N.P.
RS1	Cotton	N.A.	N.A.	N.P.	N.P.
RS2	Cotton	N.A.	N.A.	N.P.	N.P.
9S1	Cotton	1550 1250 950 850 600 450 250	1450 (1250 1100) 650	N.P.	3050 1450 750
Dh1	Digeria	1300	N.A.	N.P.	N.P.
Eup	Euphorbia	N.A.	N.A.	N.P.	N.P.
Tri	Tribularia	650 550	3150	N.P.	N.P.

Supplemental Table 2. Percent nucleotide identities between full-length DNA-A sequences analyzed

	ACMV	10c-RCA-A2-F	FJ179370	10a1-RCA-A1-F	EU589392	AM501481	3d-RCA-A1-F	GQ220850	DQ191160	10c-RCA-A1-F	FJ159262	AM421522	L2-RCA-A1-F	L4K-RCA-A1-F	AJ507777	9c-RCA-A2-F	8-RCA-A1-F	9b-RCA- A1-F	9c-RCA-A1-F	Ku-RCA-A1-F	FJ514798	AM286434	AY939926	4-RCA-A1-F	EF408037
ACMV		70	70	71	71	72	72	72	72	70	71	72	72	72	71	71	71	71	71	71	72	69	69	68	69
10c-RCA-A2-F			96	88	88	85	85	85	86	82	80	79	79	79	72	72	72	72	72	72	76	74	74	73	72
FJ179370				87	87	84	85	85	85	82	80	78	78	78	72	71	72	72	71	72	75	72	73	72	72
10a1-RCA-A1-F					99	84	84	84	84	80	80	78	78	78	73	72	73	73	73	73	80	73	73	74	72
EU589392						83	84	83	84	80	80	78	78	78	72	72	72	73	72	72	79	73	73	73	71
AM501481							100	99	94	82	84	84	84	84	74	73	74	74	73	74	79	73	73	72	72
3d-RCA-A1-F								100	94	82	84	84	84	84	74	73	74	74	73	74	79	73	73	72	72
GQ220850									94	82	84	84	84	84	74	73	74	74	73	74	79	73	73	72	72
DQ191160										83	84	87	87	87	72	72	72	73	72	73	77	73	73	73	72
10c-RCA-A1-F											89	78	78	78	77	77	77	77	77	77	75	72	72	72	70
FJ159262												80	80	80	77	77	76	77	77	77	75	71	72	71	70
AM421522													99	99	76	76	76	76	76	76	81	73	74	73	71
L2-RCA-A1-F														100	76	76	76	76	76	76	81	73	74	73	72
L4K-RCA-A1-F															76	76	76	76	76	76	81	73	74	73	72
AJ507777																96	95	96	96	96	81	71	72	71	75
9c-RCA-A2-F																	95	95	96	96	81	72	72	71	74
8-RCA-A1-F																		96	97	97	81	72	72	71	75
9b-RCA- A1-F																			98	98	81	72	72	72	75
9c-RCA-A1-F																				98	81	72	72	71	75
Ku-RCA-A1-F																					81	72	72	71	75
FJ514798																						74	75	75	73
AM286434																							96	90	70
AY939926																								91	71
4-RCA-A1-F																									71

Supplemental Table 3. Percent nucleotide identities between partial DNA-A sequences analyzed.

	10a1-RCA-A1-F	EU589392	10a1-RCA-A6-r	10c-RCA-A2-F	FJ176236	FJ179370	1b PCR A1	AM421522	L2-RCA-A1-F	L4K-RCA-A1-F	LG-RCA-A1-p	L4K PCR A1	Lc PCR A1	3d PCR A1	GQ220850	AM501481	3d-RCA-A1-F	DQ191160	3b PCR A1	6 PCR A1	FJ514798	Cgc PCR A1	TLU38239	4 PCR A1	4-RCA-A1-F	5 PCR A1	AM286434	AY939926	7 PCR A1	EF408037	ACMV	10a1-RCA-A4-p	10c-RCA-A1-F	FJ159262	8 PCR A1	8-RCA-A1-F	9c-RCA-A1-F	9a PCR A1	9b PCR A1	9b-RCA- A1-F	AJ507777	9c PCR A1	9c-RCA-A2-F	Ku-RCA-A1-F	Ku PCR A1
10a1-RCA-A1-F		99	96	96	96	96	79	79	79	79	79	79	79	85	85	85	85	85	84	80	80	81	81	76	76	76	75	76	78	78	72	73	73	76	74	74	73	74	74	74	74	74	74	74	74
EU589392			96	96	96	96	79	79	79	79	79	79	79	84	84	84	84	85	84	80	80	82	81	76	76	76	76	76	77	78	72	73	73	76	74	74	74	74	74	74	74	74	74	74	74	74
10a1-RCA-A6-r				100	99	99	80	80	80	80	80	80	80	86	86	86	86	86	85	80	80	82	81	77	77	77	76	77	78	78	71	74	74	76	75	75	74	74	74	74	74	74	74	74	74	74
10c-RCA-A2-F					99	99	80	80	80	80	80	80	80	86	86	86	86	86	85	80	80	82	81	77	77	77	76	77	78	78	71	74	74	76	75	75	74	74	74	74	74	74	74	74	74	74	74
FJ176236						99	79	79	79	79	79	79	79	86	86	85	86	86	84	80	80	82	81	77	77	77	76	76	78	78	71	74	74	76	75	75	74	74	75	75	74	75	75	75	75	75	75	75
FJ179370							79	79	79	79	79	79	79	85	85	85	85	86	84	79	80	81	81	76	76	76	76	76	78	78	71	74	74	77	75	75	74	74	74	74	74	74	74	74	74	74	74	74	74
1b PCR A1								100	100	100	100	100	100	90	90	90	90	90	91	88	89	86	85	77	77	78	77	78	82	82	73	73	73	75	76	76	77	76	77	77	76	76	76	77	77
AM421522									100	100	100	100	100	90	90	90	90	90	91	88	89	86	85	77	77	78	77	78	82	82	74	73	73	75	76	76	77	76	77	77	76	76	76	76	77	77
L2-RCA-A1-F										100	100	100	100	90	90	90	90	90	91	88	89	86	85	77	77	78	77	78	82	82	74	73	73	75	76	76	77	76	77	77	76	76	76	77	77
L4K-RCA-A1-F											100	100	100	90	90	90	90	90	91	88	89	86	85	77	77	78	77	78	82	82	74	73	73	75	76	76	77	76	77	77	76	76	76	77	77
LG-RCA-A1-p												100	100	90	90	90	90	90	91	88	89	86	85	77	77	78	77	78	82	82	74	73	73	75	76	76	77	76	77	77	76	76	76	77	77
L4K PCR A1													100	90	90	90	90	90	91	88	89	86	85	77	77	78	77	78	82	82	74	73	73	75	76	76	77	76	77	77	76	76	76	77	77
Lc PCR A1														90	90	90	90	90	91	88	89	86	85	77	77	78	77	78	82	82	74	73	73	75	76	76	77	76	77	77	76	76	76	77	77
3d PCR A1															100	100	100	100	98	82	83	82	82	76	76	77	76	77	80	80	73	71	71	74	72	72	72	73	72	72	72	73	73	73	73
GQ220850																100	100	100	98	82	83	82	82	76	76	76	76	77	80	80	73	71	71	74	72	72	72	73	72	72	72	73	73	73	73
AM501481																	100	100	98	82	83	82	82	76	76	76	76	77	80	80	73	71	71	74	72	72	72	72	72	72	72	72	73	73	73
3d-RCA-A1-F																		100	98	82	83	82	82	76	76	77	76	77	80	80	73	71	71	74	72	72	72	73	72	72	72	73	73	73	73
DQ191160																			99	82	84	83	82	76	76	77	76	77	80	80	73	71	71	74	72	72	73	73	73	72	72	73	73	73	73
3b PCR A1																				82	83	82	82	76	76	77	76	77	80	80	73	72	72	74	73	73	73	73	73	73	73	73	73	74	74
6 PCR A1																					96	87	87	79	79	80	79	79	80	80	74	73	73	75	76	76	76	76	77	77	76	77	77	77	77
FJ514798																						86	86	79	79	80	79	79	81	82	74	72	72	76	77	77	77	77	77	77	77	77	77	77	77
Cgc PCR A1																							99	81	82	82	82	81	81	71	74	74	75	76	76	76	76	76	76	76	76	76	77	77
TLU38239																								81	81	82	81	82	80	80	71	74	74	76	76	76	76	76	77	77	76	76	76	77	77
4 PCR A1																									100	98	98	98	79	79	69	73	73	73	74	74	74	74	74	74	74	74	74	74	74
4-RCA-A1-F																										98	98	97	79	79	69	73	73	73	74	74	74	74	74	74	74	74	74	74	74
5 PCR A1																											99	98	79	79	69	73	73	74	74	74	74	74	75	75	74	74	74	74	74
AM286434																												97	79	79	68	73	73	73	74	74	74	74	74	74	74	74	74	74	74
AY939926																													80	79	69	73	73	73	74	74	74	74	74	74	74	74	74	74	74
7 PCR A1																														100	72	72	72	73	73	73	74	74	74	74	73	74	74	73	73
EF408037																															72	72	72	73	73	73	74	74	74	74	73	74	74	74	74
ACMV																																69	69	70	70	70	70	70	70	70	70	70	70	70	70
10a1-RCA-A4-p																																	100	92	91	91	91	91	91	91	91	91	91	90	90
10c-RCA-A1-F																																		92	91	91	91	91	91	91	91	91	91	90	90
FJ159262																																			96	96	96	96	96	96	96	96	96	96	96
8 PCR A1																																				100	98	98	99	99	98	99	99	98	98
8-RCA-A1-F																																					98	98	99	99	98	99	99	98	98
9c-RCA-A1-F																																						98	99	99	99	99	99	98	98
9a PCR A1																																							99	99	99	99	99	99	99
9b PCR A1																																								100	99	100	100	99	99
9b-RCA- A1-F																																									99	100	100	99	99
AJ507777																																										99	99	99	99
9c PCR A1																																											100	99	99
9c-RCA-A2-F																																												99	99
Ku-RCA-A1-F																																													100

Supplemental Table 4. Percent nucleotide identities between partial betasatellite sequences analyzed.

	DQ364230	AY795604	HS1 PCR B1	HS2 PCR B1	AJ316035	L5K PCR B1	3d PCR B1	AY820772	EU384591	HYVMDaSat	AM084379	L4K PCR B1	RS1 PCR B1	L2-RCA-b1-F	L2 PCR B1	LG PCR B1	LG-RCA-b1-r	RS2-PCR B1
DQ364230		98	99	99	96	94	85	85	78	56	96	93	92	93	93	93	93	92
AY795604			98	98	97	94	84	84	78	56	94	91	90	91	91	91	91	91
HS1 PCR B1				100	96	94	85	85	78	56	95	92	91	92	92	92	92	92
HS2 PCR B1					96	94	85	85	78	55	94	92	91	92	92	92	92	91
AJ316035						97	83	83	78	55	93	90	89	90	90	90	90	90
L5K PCR B1							84	84	80	56	91	89	88	89	89	89	89	89
3d PCR B1								100	85	54	81	84	83	84	84	84	84	84
AY820772									85	54	81	84	83	84	84	84	84	84
EU384591										54	76	77	77	77	77	77	77	77
HYVMDaSat											56	56	57	57	57	56	56	57
AM084379												96	95	96	96	96	96	95
L4K PCR B1													99	100	100	100	100	99
RS1 PCR B1														98	98	98	98	98
L2-RCA-b1-F															100	100	100	100
L2 PCR B1																100	100	100
LG PCR B1																	100	100
LG-RCA-b1-r																		100

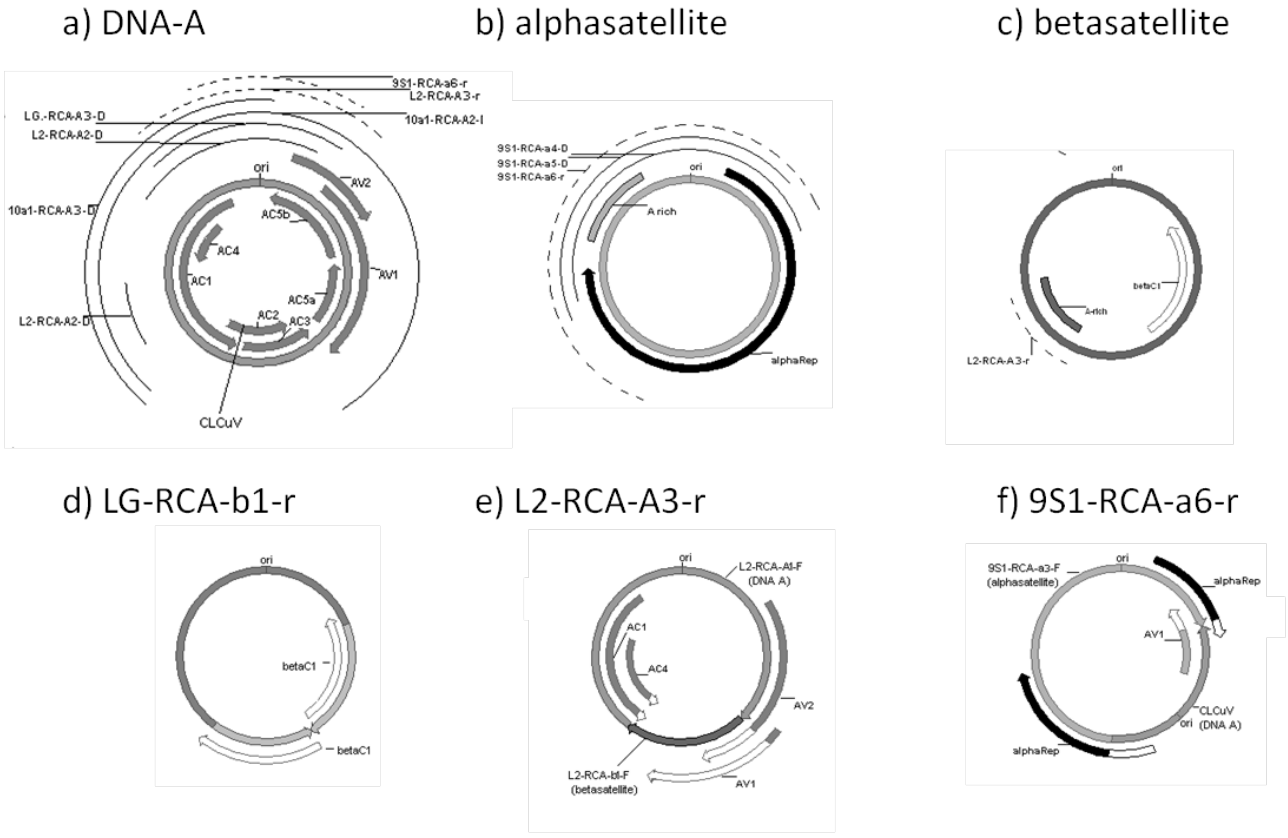
Supplemental Table 5. Percent nucleotide identities between full-length alphasatellite sequences analyzed.

	AJ132345	9S1-RCA-a1-F	LG-RCA-a1-F	AM711116	9S1-RCA-a3-F	EU384637	9S1-RCA-a2-F	EU384611	AJ512948	AJ512958	10a1-RCA-a1-F	10c-RCA-a1-F	10b-RCA-a1-F	AJ416153	FJ956707	9c-RCA-a2-F	FN436007	FN436008	HM163578	CFDV	9a-RCA-a1-F	EU384652	AM711115	FN678901
AJ132345		90	90	83	78	78	78	78	79	79	80	80	80	47	46	47	46	46	48	42	57	59	75	75
9S1-RCA-a1-F			100	82	76	77	78	78	78	79	81	81	81	47	46	48	45	46	47	44	57	59	76	75
LG-RCA-a1-F				82	76	77	78	78	78	79	81	81	81	47	46	48	45	46	47	44	57	59	76	75
AM711116					84	83	84	84	80	80	85	85	85	48	48	47	46	47	47	43	57	59	75	76
9S1-RCA-a3-F						97	95	93	78	78	78	78	78	48	48	48	46	46	47	44	59	61	74	75
EU384637							95	94	78	78	78	78	78	48	48	49	47	47	48	45	59	62	74	76
9S1-RCA-a2-F								98	78	78	79	79	79	48	48	48	46	47	47	45	59	61	74	75
EU384611									77	77	79	79	79	48	48	48	46	46	47	44	58	60	74	75
AJ512948										96	86	86	86	49	48	48	46	45	46	46	58	60	78	86
AJ512958											86	86	86	49	48	48	46	45	46	45	58	59	79	86
10a1-RCA-a1-F												100	100	48	48	48	46	45	46	44	57	59	76	80
10c-RCA-a1-F													100	48	48	48	46	45	46	44	57	59	76	80
10b-RCA-a1-F														48	48	48	46	45	46	44	57	59	76	80
AJ416153															93	55	53	53	53	45	45	47	46	48
FJ956707																55	53	52	52	44	44	46	46	48
9c-RCA-a2-F																	68	66	68	44	45	46	47	49
FN436007																		76	73	41	43	44	46	47
FN436008																			72	42	44	44	44	47
HM163578																				40	45	46	46	47
CFDV																					47	47	45	46
9a-RCA-a1-F																						92	57	57
EU384652																							60	60
AM711115																								81

Supplemental Table 6. The *P*-values measured by the different methods used by RDP 3.27 for the recombination event at nucleotide positions 1108 and 1458 resulting in molecule 10c-RCA-A1-F, shown in Fig. 8.

Method	*P-value*
RDP	1.68×10^{-11}
GENECONV	4.39×10^{-04}
Bootscan	1.70×10^{-09}
Maxchi	4.03×10^{-09}
Chimaera	4.33×10^{-10}
SiScan	1.58×10^{-11}
PhylPro	Ns
LARD	Ns
3Seq	1.78×10^{-09}

Suppl. Fig. 1



Suppl Fig. 1 Genome maps of unusual or novel circular DNAs. (a) Begomoviral full-length DNA-A showing the positions of the predicted genes (shown by arrows) encoded in both virion and complementary sense. Externally, the positions on the complete genome of the defective molecules are shown by full lines, and the segments in recombinants by dashed lines. (b) Begomoviral full-length alphasatellite DNA showing the positions of the alphaRep gene and the A-rich region. Externally, the positions on the complete genome of the defective molecules are shown by full lines, and the segments in recombinants by dashed lines. (c) Begomoviral full-length betasatellite showing the positions of the betaC1 gene and the A-rich region. Externally the position on the complete molecule of the segment present in recombinant L2-RCA-A3-r is shown. (d) Molecule LG-RCA-b1-r, a recombinant betasatellite molecule. The sequence in light gray is an inverted repeat. The upper inverted betaC1 ORF is truncated. (e) Molecule L2-RCA-A3-r, a recombinant between the DNA-A molecule L2-RCA-A1-F and the betasatellite molecule L2-RCA-b1-F. Four DNA-A ORFs are shown, which terminate inside the betasatellite-derived sequence. The positions on the DNA-A and on the betasatellite of the parental sequences are shown in figure 7a and7b. (f) Molecule 9S1-RCA-a6-r, a recombinant between the alphasatellite molecule 9S1-RCA-a3-F and a CLCuV DNA-A molecule (whose sequence is inverted). The two origins of replication (ori) are shown, along with the ORFs, whose partial sequence is identical to the alphaRep and the AV1 genes. The positions on the DNA-A and on the alphasatellite DNA of the parental sequences are shown in figure 7a and 7c. Maps not shown to scale.
Please move to SOM