

Types of variation in DNA-A among isolates of East African cassava mosaic virus from Kenya, Malawi and Tanzania

Xueping Zhou,^{1,2} David J. Robinson¹ and Bryan D. Harrison¹

¹ Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK

² Institute of Biotechnology, Zhejiang Agricultural University, Hangzhou 310029, China

Complete nucleotide sequences of the DNA-A-like molecules of three East African cassava mosaic virus (EACMV) isolates from Kenya (-K, 2801 nt) and Malawi (-MH and -MK, both 2804 nt) were determined. These sequences were compared with that published for a Tanzanian isolate (-T, 2801 nt) and the partial sequence of a third Malawian isolate. Intergenic region sequences of all isolates, and deduced amino acid sequences of their AC1 (Rep) proteins, each formed a tightly related cluster that was distinct from the comparable components of other begomoviruses. Other complementary-sense genes (AC2, AC3, AC4) differed between EACMV isolates in a way consistent with the accumulation of point mutations. In contrast, virus-sense genes (CP, AV2) of isolates -MH and -MK differed (substantially for AV2) from those of other EACMV isolates but somewhat resembled those of tomato yellow leaf curl virus – Israel, suggesting they had been acquired by recombination with an unidentified begomovirus.

East African cassava mosaic virus (EACMV), African cassava mosaic virus (ACMV) and Indian cassava mosaic virus (ICMV) are the three whitefly-transmitted geminiviruses (genus *Begomovirus*) known to be associated with cassava mosaic disease in Africa and the Indian subcontinent. They are distinguished by the nucleotide sequences of their genomic DNA molecules (Stanley & Gay, 1983; Hong *et al.*, 1993; Zhou *et al.*, 1997) and by their reactions with a panel of monoclonal antibodies (Harrison & Robinson, 1988; Swanson & Harrison, 1994). EACMV occurs in Kenya, Madagascar, Malawi, Tanzania and Zimbabwe (Swanson & Harrison, 1994) and probably also in Zambia (Ogbe *et al.*, 1997) and Mozambique (M. M. Swanson & B. D. Harrison, unpublished

Author for correspondence: Bryan D. Harrison.

Fax +44 1382 562426. e-mail djrobi@scri.sari.ac.uk

The EMBL accession numbers of the sequences reported in this paper are AJ006458–AJ006460.

results). In addition, a recombinant virus (UgV), the DNA-A of which consists of a large part of the coat protein gene of ACMV inserted in an otherwise EACMV-like sequence, was discovered infecting cassava in Uganda (Zhou *et al.*, 1997). Recombination has apparently also played an important part in the evolution of the begomoviruses causing the current epidemic of cotton leaf curl disease in Pakistan (Zhou *et al.*, 1998). In this paper, we report the results of work designed to explore the extent and nature of variation in DNA-A among EACMV isolates from widely separated locations in Kenya, Malawi and Tanzania, and we describe evidence that two isolates are recombinants.

The virus isolates compared in most detail were:

EACMV-K2B, from Msabaha, coastal Kenya (= EACMV-K; Zhou *et al.*, 1997)

EACMV-MalH, from Chitedze, Malawi (= EACMV-MH)

EACMV-MalK, from Chitedze, Malawi (= EACMV-MK)

EACMV-Dar6, from Dar es Salaam, Tanzania (= EACMV-T; Zhou *et al.*, 1997)

EACMV-YG, from Rumphi, Malawi (Hong *et al.*, 1993)

ACMV-K, from Western Kenya (Stanley & Gay, 1983)

Tomato yellow leaf curl virus – Israel (= TYLCV-Is; Navot *et al.*, 1991)

UgV-C39, from Otuboi, Uganda (Zhou *et al.*, 1997)

Infected cassava plants were grown in a containment glasshouse under licence from the Scottish Office Agriculture, Environment and Fisheries Department. Viral DNA was extracted, amplified by PCR and cloned as described previously (Zhou *et al.*, 1997). The complete sequences of molecules equivalent to DNA-A were determined for EACMV-K, EACMV-MH and EACMV-MK by the procedures of Zhou *et al.* (1997) except that two additional primers were used for the two Malawi isolates [positions in the complete nucleotide sequence are numbered from the origin of replication (*ori*; Laufs *et al.*, 1995)]:

MAL-L3/F 5' TACGCATGCCTCTAATCCAG 3' (nt 1038–1057)

MAL-L1/R 5' TTCCGCCACAACCTTATGTA 3' (complementary to nt 2024–2005)

Each viral DNA-A was covered by three clones, with overlaps of 108, 63 and 51 nt at the three junctions. In every instance the overlapping sequences were identical. Each

Table 1. Percentage nucleotide or amino acid sequence identities of EACMV isolates and selected other begomoviruses**(a) Nucleotide sequences**

Virus isolate	EACMV-T	EACMV-K	UgV-C39	EACMV-YG	EACMV-MH	EACMV-MK	ACMV-K
	Complete DNA-A compared:						
EACMV-T	–	94	91	ND	86	86	68
EACMV-K	90	–	92	ND	87	87	68
UgV-C39	95	92	–	ND	86	86	72
EACMV-YG	80	79	81	–	ND	ND	ND
EACMV-MH	84	89	85	77	–	98	70
EACMV-MK	84	90	86	78	96	–	69
ACMV-K	54	53	55	57	52	52	–
	Intergenic region compared:						

(b) Amino acid sequences of AV2 protein

Virus isolate	EACMV-T	EACMV-K	UgV-C39	EACMV-YG	EACMV-MH	EACMV-MK	ACMV-K
EACMV-T	–	97	95	93	62	62	60
EACMV-MH	62	63	61	61	–	100	76
TYLCV-IS	63	64	63	65	87	87	73

–, Comparison with self.

ND, Not determined.

nucleotide was determined at least twice. The three complete sequences (EACMV-K, 2801 nt; EACMV-MH, 2804 nt; EACMV-MK, 2804 nt) were compared with the partial sequence of EACMV-YG (Hong *et al.*, 1993; AJ006461), and the published sequences of DNA-A of EACMV-T (2801 nt; Z83256), UgV-C39 (2799 nt; Z83257), ACMV-K (2779 nt; J02057) and TYLCV-Is (2787 nt; X15656).

In length and percentage sequence identity (Table 1a), the EACMV sequences were of two types: (1) EACMV-T and EACMV-K, and (2) EACMV-MH and EACMV-MK, with 94–98% sequence identity within a type but only 86–87% between types. UgV-C39 was more like the first type [91–92%, rising to 96% when its recombinant coat protein (*CP*) gene was excluded from the calculation] than the second type (86%). All the EACMV isolates are almost equally different from ACMV-K (68–70% sequence identity; Table 1a), the extent of the difference being typical of that between different begomovirus species. The difference in length of the DNA-A sequences of the two types of EACMV is attributable to an in-frame insertion/deletion of 3 nt in the *CP* gene, and the smaller size of the UgV-C39 sequence is accounted for by a deletion of 2 nt in the intergenic region.

Comparisons of sequences of the intergenic region of begomoviruses probably give the most sensitive indication of the extent of viral similarities and differences (Padidam *et al.*, 1995). EACMV isolates of types 1 and 2 have intergenic regions of the same length (330 nt) and that of UgV-C39

differs only slightly (328 nt) by lacking the two residues equivalent to nt 197 and 198 in the other sequences. In contrast, the intergenic region of EACMV-YG contains only 317 nt, and that of ACMV-K contains 308 nt. The percentage sequence identities among the intergenic regions of these isolates follow a similar pattern to those for the complete DNA-A sequences but are mostly lower, with EACMV-YG differing from EACMV isolates of types 1 and 2 to a greater extent than they differ from one another (Table 1a). ACMV-K differs greatly (only 52–57% identity) from all the other isolates.

All the intergenic region sequences contain various features characteristic of begomoviruses: a putative stem-loop structure, with the nonanucleotide TAATATTAC in the loop; a TATA motif at nt 67–70 of the intergenic sequence of types 1 and 2 isolates; and repeated sequences to the 5' side of the TATA motif (Argüello-Astorga *et al.*, 1994). These repeats are GGGGRCA for EACMV types 1 and 2 isolates but AAT-TGGAGACA for EACMV-YG. Interestingly, the EACMV-YG repeat is the same as that in ACMV-K and cotton leaf curl virus – Pakistan (CLCuV-PK) type 802a, and the GGGGRCA repeat is also found in CLCuV-PK type 26 (Zhou *et al.*, 1998). Evidently these motifs recur in somewhat distantly related viruses, although they can differ in closely related ones (Zhou *et al.*, 1998). These similarities with ACMV-K and CLCuV-PK extend only a little or not at all to the sequences surrounding the repeats.

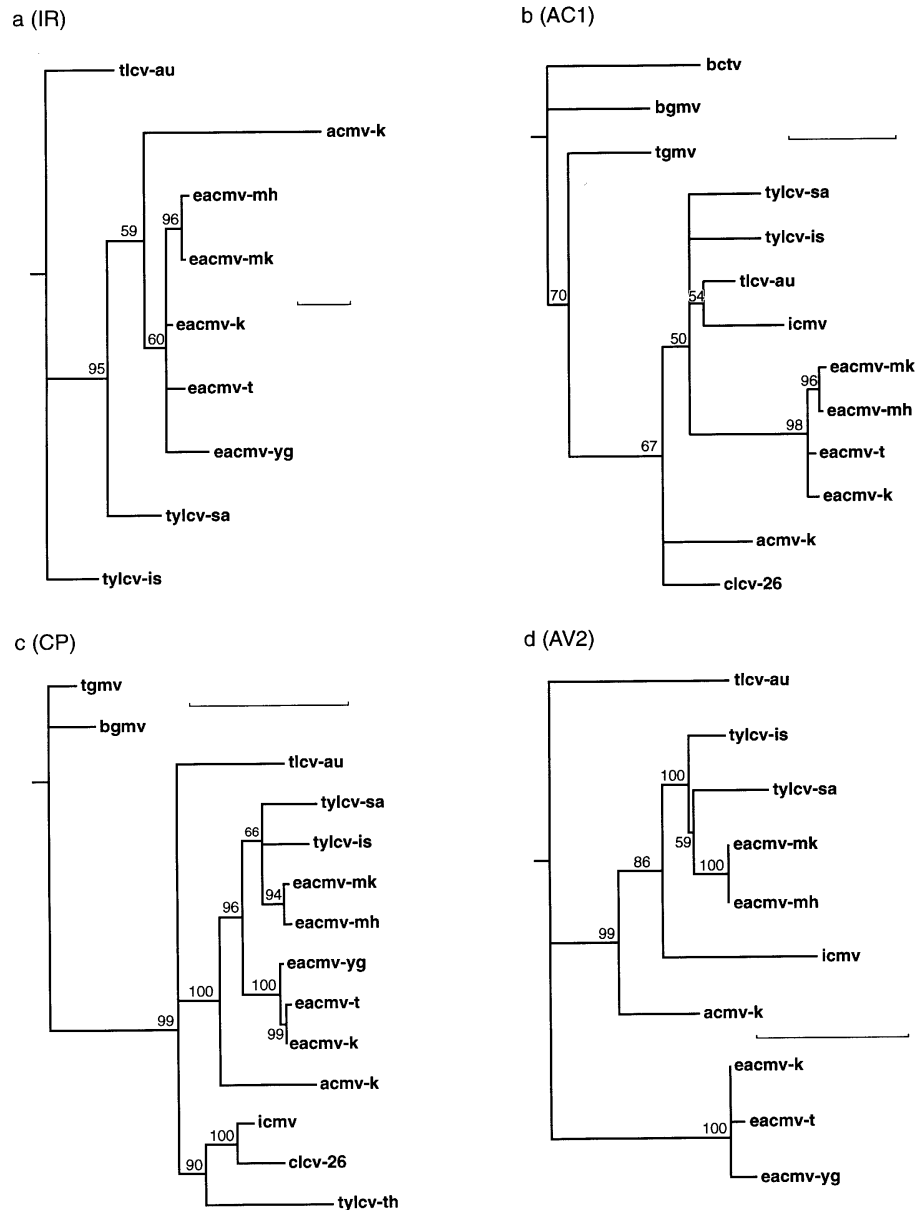


Fig. 1. Phylogenetic trees obtained using the PUZZLE program from alignments of geminivirus sequences. Numbers above each branch are the percentage support for that branch. Scale bars indicate the horizontal distances equivalent to 0.2 replacements per position. (a) Nucleotide sequences of the intergenic region of EACMV and other begomovirus isolates from the African–Mediterranean region, with tomato leaf curl virus, Australian isolate as an outgroup. (b) Deduced amino acid sequences of the AC1 (Rep) protein of EACMV isolates, eight other begomoviruses from five continents and the leafhopper-transmitted beet curly top virus. (c) Deduced amino acid sequences of the CP of EACMV isolates and nine other begomoviruses from five continents. (d) Deduced amino acid sequences of the AV2 protein of EACMV isolates and five other begomoviruses from the Eastern Hemisphere. Abbreviations for viruses other than EACMV represent: acmv-k, African cassava mosaic virus, Kenyan isolate (database accession number J02057); bgmv, bean golden mosaic virus (M10070); bctv, beet curly top virus (X04144); clcv-26, cotton leaf curl virus – Pakistan, type 26 (AJ002458); icmv, Indian cassava mosaic virus (Z24758); tgm, tomato golden mosaic virus (K02029); tylcv-au, tomato leaf curl virus, Australian isolate (S53251); tylcv-is, tomato yellow leaf curl virus, Israeli isolate (X15656); tylcv-sa, tomato yellow leaf curl virus, Sardinian isolate (X61153); tylcv-th, tomato yellow leaf curl virus, Thailand isolate (M59839).

The relationships between the intergenic regions of these and some other begomoviruses are shown in a phylogenetic tree (Fig. 1a). All the EACMV isolates form a tight cluster which is well separated from ACMV-K and from two tomato

begomoviruses from the Mediterranean region. In this analysis, TLCV-Au is used as an outgroup but other begomoviruses were omitted because the many nucleotide differences in their intergenic regions made sequence alignments highly arbitrary.

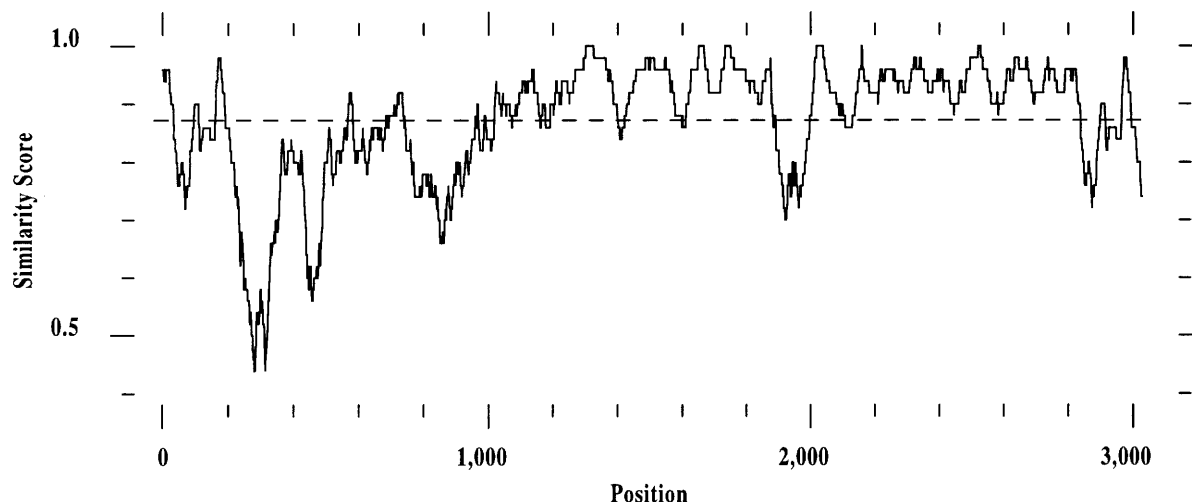


Fig. 2. PLOTSIMILARITY diagram (Anon., 1994) comparing the DNA-A sequences of EACMV-K and EACMV-MK. Nucleotides are numbered from *ori*, and the first 225 nt are shown at each end of the sequence. The horizontal broken line represents the mean difference between the two sequences. Scanning window = 50 nt.

Fig. 1(b) shows a comparable phylogenetic tree based on deduced sequences of AC1 proteins (replication-associated protein, Rep), but including a wider range of geminiviruses. Again the EACMV isolates form a tight cluster which is well separated from other begomoviruses from the Eastern Hemisphere and is even more distant from viruses from the Western Hemisphere (bean golden mosaic virus, tomato golden mosaic virus).

Previous analyses (Hong & Harrison, 1995) showed that a phylogenetic tree of the CPs of begomoviruses had three main branches, corresponding to the geographical region of occurrence of the viruses: America, the African-Mediterranean area or Asia. EACMV isolates all are included in the African-Mediterranean branch but they are distributed between two sub-branches in such a way that type 2 isolates are closer to the begomoviruses from tomato than to type 1 isolates (Fig. 1c). CPs of EACMV isolates within a type had 98–99% amino acid sequence identity whereas isolates of different types had 87–88% identity. The CP sequence identity of EACMV-MH to TYLCV-Is was 90%.

Comparable relationships for the AV2 protein showed an even more marked separation of the EACMV isolates into the two clusters distinguished on the basis of CP sequences (Fig. 1d). The AV2 proteins of EACMV-MH and EACMV-MK were much more like that of TYLCV-Is (87% sequence identity) than those of EACMV type 1 isolates (only 61–63%; Table 1b).

Because of the differences in the CP sequences, and especially the AV2 protein sequences, between the two types of EACMV, their inclusion in the same viral species needs to be justified. We consider that it is warranted by the similarities of their deduced Rep (94–98% sequence identity), AC2 (89–99%), AC3 (90–96%) and AC4 (87–95%) proteins, and by the grouping of their intergenic region sequences in a tight cluster

well separated from other viruses. Furthermore, isolates of both types have epitope profiles typical of EACMV (Swanson & Harrison, 1994) when tested against a standard panel of geminivirus monoclonal antibodies that differentiated EACMV from all but one (TYLCV-Senegal) of 40 other begomoviruses (results not shown).

Comparison of deduced sequences of the AC2, AC3 and AC4 proteins gives sequence identities of 87–92% between any two of the EACMV isolates except that the figures for comparison of EACMV-MH and EACMV-MK are 95–99%. Inspection of the nucleotide sequences of the AC2 and AC3 genes showed that substitutions are scattered along them, and that several do not change the amino acid encoded. This pattern of variation is therefore typical of that produced as the result of the accumulation of point mutations in different populations of a virus by genetic drift.

In contrast, comparison of the AV2 plus CP sequences indicates that their genes must have different evolutionary origins in the two types of EACMV. The possibility that the difference is an artifact caused by clones having been derived from DNA-A of two different begomoviruses coinfecting any of the source plants of the EACMV isolates seems remote. The clone containing the AV2 and CP genes of each virus had overlaps of 51 and 108 nt with the adjacent clones, and the overlaps had identical sequences. Moreover, the clone containing the AV2 and CP genes extended into the AC1 and AC3 genes, revealing only trivial differences between the two types of EACMV in these regions. Thus the main difference between the two types of EACMV is in their AV2 and, to a lesser extent, their CP genes. We therefore consider a much more plausible explanation for this difference is that, in one type of EACMV, the AV2 and CP genes were acquired by recombination between an original form of EACMV and another begomovirus. A comparable situation, in which virus-sense

genes apparently can have a different evolutionary origin from complementary-sense genes, and *ori* seems to be a recombination hot spot, occurs among begomoviruses infecting cotton and okra in Pakistan (Zhou *et al.*, 1998). If the same is true for EACMV, the extent of the sequence dissimilarity between type 1 and type 2 isolates in the intergenic region would be expected to be different on either side of *ori*. Inspection of the intergenic region sequences shows that, to the 5' side of *ori*, coordinated differences between EACMV-MH and -MK on the one hand and EACMV-K, -T and -YG on the other hand occur at only 3 positions (out of 158). However, on the 3' side of *ori*, usually a more conserved region, coordinated differences occur at 13 positions out of 172. Our analysis therefore suggests that recombination has played a key role in the divergence of the two types of EACMV and that *ori* may be a recombination point.

Evidence on this matter was also sought by comparing all parts of the EACMV-MK and EACMV-K DNA-A sequences, with the aid of a PLOTSIMILARITY diagram (Fig. 2). When interpreting such plots, it must be remembered that some parts of begomovirus DNA-A are much more prone to mutation than others which, in contrast, may be conserved in distinct begomovirus species. In Fig. 2, three zones of dissimilarity, in which similarity values drop below 0.7, can be seen. These are not found by comparing EACMV-MK with EACMV-MH (at least 0.9 throughout) or EACMV-K with EACMV-T [> 0.8 throughout except for part of the 5' half of the intergenic region (> 0.7)]. The three zones of dissimilarity are located in (1) the *AV2* gene (about nt 200–350), (2) the overlapping parts of the *AV2* and *CP* genes, extending into the *CP* gene (nt 425–650), and (3) part of the 3' half of the *CP* gene (nt 750–1000). A subsidiary zone of dissimilarity lies a little to the 3' side of *ori* (nt 20–100). These four zones are separated by relatively similar sequences which mostly tend to be conserved in different begomovirus species. This interpretation of Fig. 2 is therefore entirely compatible with our suggestion that a progenitor of either type 1 or type 2 EACMV isolates had acquired from another begomovirus a sequence stretching from near nt 1 to a point within the range of nt 1050–1150.

In addition to the four zones discussed above, a fifth zone of difference between EACMV-MK and EACMV-K was detected at about nt 1920–2000 (Fig. 2). This zone also occurred when EACMV-K was compared with EACMV-MH, or EACMV-T with EACMV-MH or EACMV-MK. We have no further evidence on the origin of this difference.

In studies on the recombinant begomovirus UgV, its parents (EACMV and ACMV) could be identified with confidence (Zhou *et al.*, 1997). Our present results show that UgV is more closely related to type 1 than to type 2 isolates of EACMV, indicating that its EACMV parent was probably a type 1 isolate. In the two types of EACMV, the respective origins of the sequence to the 3' side of *ori* are less clear. Those found in type 1 occur in isolates from Kenya, Malawi and Tanzania (and for the *AV2* gene, Uganda). These sequences

differ considerably from those in the comparable region of other begomoviruses for which data are available, which may either indicate the distinctiveness of EACMV or simply reflect the lack of sequence data for its closest relatives. In contrast, the sequences of EACMV-MH and -MK to the 3' side of *ori* have obvious affinities with that of TYLCV-Is. However, TYLCV-Is is unlikely itself to have been the source of these sequences because its *CP* gene (usually the most strongly conserved part of the begomovirus genome) is not similar enough to the *CP* genes of EACMV-MH and -MK. Moreover, the TYLCV-Is intergenic region sequence to the 3' side of *ori* is unlike the comparable sequences of EACMV-MH and -MK. Our data are therefore compatible with the view that an EACMV type 1 isolate is the older form, and that type 2 isolates have been derived by recombination with a begomovirus that has affinities to TYLCV-Is; however, other possibilities are not ruled out.

We thank Mr J. P. Kamau (Kenya) and Dr P. Subrahmanyam (Malawi) for sending infected cassava cuttings. X. Zhou held a Sino-Scottish Fellowship in Tropical Plant Virology. SCRI is grant-aided by the Scottish Office Agriculture, Environment and Fisheries Department.

References

- Anon. (1994). Program Manual for the Wisconsin Package, Version 8. Madison: Genetics Computer Group.
- Argüello-Astorga, G., Herrera-Estrella, L. & Rivera-Bustamente, R. (1994). Experimental and theoretical definition of geminivirus origin of replication. *Plant Molecular Biology* **26**, 553–556.
- Harrison, B. D. & Robinson, D. J. (1988). Molecular variation in vector-borne plant viruses: epidemiological significance. *Philosophical Transactions of the Royal Society of London Series B* **321**, 447–462.
- Hong, Y. G. & Harrison, B. D. (1995). Nucleotide sequences from tomato leaf curl viruses from different countries: evidence for three geographically separate branches in evolution of the coat protein of whitefly-transmitted geminiviruses. *Journal of General Virology* **76**, 2043–2049.
- Hong, Y. G., Robinson, D. J. & Harrison, B. D. (1993). Nucleotide sequence evidence for the occurrence of three distinct whitefly-transmitted geminiviruses in cassava. *Journal of General Virology* **74**, 2437–2443.
- Laufs, J., Traut, W., Heyraud, F., Matzeit, V., Rogers, S. G., Schell, J. & Gronenborn, B. (1995). *In vitro* cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. *Proceedings of the National Academy of Sciences, USA* **92**, 3879–3883.
- Navot, N., Pichersky, E., Zeidan, M., Zamir, D. & Czosnek, H. (1991). Tomato yellow leaf curl virus: a whitefly-transmitted geminivirus with a single genomic molecule. *Virology* **185**, 151–161.
- Ogbe, F. O., Legg, J., Raya, M. D., Muimba-Kankalongo, A., Theu, M. P., Kaitisha, G., Phiri, N. A. & Chalwe, A. (1997). Diagnostic survey of cassava mosaic viruses in Tanzania, Malawi and Zambia. *Roots* **4**, 12–15.
- Padidam, M., Beachy, R. N. & Fauquet, C. M. (1995). Classification and identification of geminiviruses using sequence comparisons. *Journal of General Virology* **76**, 249–263.
- Stanley, J. & Gay, M. R. (1983). Nucleotide sequence of cassava latent virus DNA. *Nature* **301**, 260–262.

Swanson, M. M. & Harrison, B. D. (1994). Properties, relationships and distribution of cassava mosaic geminiviruses. *Tropical Science* **34**, 15–25.

Zhou, X., Liu, Y., Calvert, L., Munoz, C., Otim-Nape, G. W., Robinson, D. J. & Harrison, B. D. (1997). Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen by interspecific recombination. *Journal of General Virology* **78**, 2101–2111.

Zhou, X., Liu, Y., Robinson, D. J. & Harrison, B. D. (1998). Four DNA-A variants among Pakistani isolates of cotton leaf curl virus and their affinities to DNA-A of geminivirus isolates from okra. *Journal of General Virology* **79**, 915–923.

Received 28 May 1998; Accepted 5 August 1998