Sequence changes in six variants of rice tungro bacilliform virus and their phylogenetic relationships

Pepito Q. Cabauatan,[1] Ulrich Melcher,[2] Koichi Ishikawa,[3] Toshihiro Omura,[3] Hiroyuki Hibino,[4] Hiroki Koganezawa[5] and Ossmat Azzam[1]

[1] Entomology and Plant Pathology Division, International Rice Research Institute, PO Box 933, 1099 Manila, Philippines

[2] Oklahoma State University, Stillwater, OK 74078, USA

[3] National Agriculture Research Center, Tsukuba 305-0856, Japan

[4] National Institute for Agroenvironmental Science, Tsukuba 305-8604, Japan

[5] Shikoku National Agricultural Experiment Station, Zentsuji, Kagawa 765-0001, Japan

The DNA of three biological variants, G1, Ic and G2, which originated from the same greenhouse isolate of rice tungro bacilliform virus (RTBV) at the International Rice Research Institute (IRRI), was cloned and sequenced. Comparison of the sequences revealed small differences in genome sizes. The variants were between 95 and 99% identical at the nucleotide and amino acid levels. Alignment of the three genome sequences with those of three published RTBV sequences (Phi-1, Phi-2 and Phi-3) revealed numerous nucleotide substitutions and some insertions and deletions. The published RTBV sequences originated from the same greenhouse isolate at IRRI 20, 11 and 9 years ago. All open reading frames (ORFs) and known functional domains were conserved across the six variants. The cysteine-rich region of ORF3 showed the greatest variation. When the six DNA sequences from IRRI were compared with that of an isolate from Malaysia (Serdang), similar changes were observed in the cysteine-rich region in addition to other nucleotide substitutions and deletions across the genome. The aligned nucleotide sequences of the IRRI variants and Serdang were used to analyse phylogenetic relationships by the bootstrapped parsimony, distance and maximum-likelihood methods. The isolates clustered in three groups: Serdang alone; Ic and G1; and Phi-1, Phi-2, Phi-3 and G2. The distribution of phylogenetically informative residues in the IRRI sequences shared with the Serdang sequence and the differing tree topologies for segments of the genome suggested that recombination, as well as substitutions and insertions or deletions, has played a role in the evolution of RTBV variants. The significance and implications of these evolutionary forces are discussed in comparison with badnaviruses and caulimoviruses.

Introduction

Rice tungro bacilliform virus (RTBV) is one of the two viruses that synergistically cause tungro disease of rice. RTBV causes the tungro symptoms, which include stunting and yellow to orange discoloration of the infected plants (Hibino *et al.*, 1978). RTBV has an 8002 bp circular double-stranded DNA genome with two discontinuities, one on each strand (Hay *et al.*, 1991; Qu *et al.*, 1991; Hibino *et al.*, 1991). RTBV has four open reading frames (ORFs), potentially capable of encoding proteins of 24, 12, 194 and 46 kDa (Hay *et al.*, 1991). ORF3 encodes a P194 polyprotein that contains four functional domains: the viral coat protein, aspartate protease, reverse transcriptase and ribonuclease H. Sequences of three RTBV variants originating from the International Rice Research Institute (IRRI) in the Philippines were reported previously and are referred to here as Phi-1 (Hay *et al.*, 1991), Phi-2 (Qu *et al.*, 1991) and Phi-3 (Kano *et al.*, 1992). In addition, the sequence of one isolate from Malaysia (Serdang; EMBL accession no. 076470) has recently become available. The published IRRI sequences differ from one another by about 100 nucleotide

Author for correspondence: Ossmat Azzam.
Fax +63 2 891 1292. e-mail O.Azzam@cgiar.org

The EMBL accession numbers of the RTBV DNA sequences reported in this paper are AF113830–AF113832.

0001-6208 © 1999 SGM

substitutions scattered throughout the genome (Hull, 1996). This finding suggested heterogeneity of the virus isolate maintained at the IRRI greenhouse. This suggestion received support when Villegas *et al.* (1997) analysed 27 full-length *Bam*HI clones from this isolate. The clones could be grouped into at least three *Pst*I and *Eco*RV RFLP types. Moreover, sequence data obtained from a small segment (nt 7681–7982) of the 27 clones revealed considerable microvariation within the IRRI isolate. Variation in the genome of RTBV was also demonstrated by Fan *et al.* (1996) in isolates from India, Bangladesh, Indonesia, Malaysia and Thailand. They showed, by restriction mapping, cross-hybridization and sequencing of the intergenic region, that RTBV isolates fall into two groups, those from the Indian subcontinent and those from the south-east Asian countries. The two groups, or strains, have a low percentage sequence identity, and a 64 base deletion in the intergenic region differentiated one group from the other.

Cabauatan *et al.* (1995) isolated four biological variants (G1, Ic, G2 and L) from the same IRRI isolate as above, based on the symptoms induced on rice varieties FK135 and TN1. Preliminary studies revealed that the four variants were genetically different. Restriction analysis of viral DNA showed that the G1 and Ic variants had identical restriction maps but differed from G2 and L in *Pst*I, *Eco*RI and *Eco*RV sites (Cabauatan *et al.*, 1998). This study confirms that the RTBV isolate maintained at IRRI is composed of a heterogeneous population of biological and genetic variants. We were interested in examining sequence changes in that population so that possible causes and effects of adaptive or evolutionary changes can be elucidated. Such information should be critical in designing effective and durable resistance to RTBV.

Here we report the cloning and sequencing of variants G1, Ic and G2 and the comparison of complete genome sequences with each other and with those of the previously determined sequences from IRRI and Malaysia. DNA sequence analysis revealed that the seven variants differed in genome size and that such differences were due to insertions and deletions in both the coding and the non-coding regions. All ORFs and known functional domains were conserved across the six variants. The cysteine-rich region of ORF3 showed the greatest variation. Phylogenetic analysis of the seven sequences showed that the IRRI isolate consisted of at least two clusters and that it was significantly different from the Malaysian isolate. Roles for nucleotide substitution and recombination in RTBV evolution were evaluated.

Methods

■ **RTBV isolates and DNA isolation.** RTBV variants G1, Ic, G2 and L, which have been maintained in the IRRI greenhouse since 1994, were propagated in rice cv. TN1 by insect transmission using *Nephottetix virescens* and the virus was purified from the infected plants as described previously (Cabauatan & Hibino, 1988; Cabauatan *et al.*, 1995). Isolate Phi-1 was sent to the John Innes Institute, Norwich, UK, as infected plants in 1987 while Phi-2 was brought to the Scripps Research Institute, La Jolla, CA, USA, as viral DNA purified from the same greenhouse isolate some time in 1987 and 1989. Phi-3, on the other hand, was brought to the National Agriculture Research Center (NARC), Tsukuba, Japan, more than 20 years ago and maintained on rice cv. TN1 in an air-conditioned greenhouse by successive transfers using *N. virescens*. It was included for comparison in the initial experiments to predict endonuclease maps and to differentiate the variants by PCR. All six isolates described were derived from the same IRRI greenhouse isolate. Viral DNA was extracted from purified suspensions by SDS–phenol extraction after pre-treatment with protease K (2·5 mg/ml) for 1 h, followed by ethanol precipitation.

■ **Restriction endonuclease mapping and PCR differentiation of RTBV variants.** Viral DNA from G1, Ic, G2, L and Phi-3 was cleaved with *Bam*HI, *Bst*XI, *Eco*RI, *Eco*RV and *Pst*I and electrophoresed in agarose gels. The cleavage sites were deduced as described previously (Hibino *et al.*, 1991). Restriction endonuclease cleavage sites for G1, Ic and G2, which were deduced by enzymatic cutting pattern analysis, were confirmed by nucleotide sequencing.

On the basis of a comparison between G2, Phi-2 and Phi-3 sequences, several primers were designed to amplify different regions of the RTBV genome by PCR (Table 1). PCR conditions were as described by Takahashi *et al.* (1993). The products were then either digested with *Eco*RI (for primer pair P3/P7) or *Eco*RI and *Eco*RV (for primer pair P5/P6), electrophoresed in 1% agarose gels and viewed under UV light after staining with ethidium bromide.

■ **Cloning and sequencing.** G2 DNA was cleaved with *Pst*I and cloned into pBluescript II KS (+) (Stratagene). Two types of clones, one with an insert of about 5 kbp and another with one of about 3 kbp, were generated. A series of unidirectionally deleted DNAs was generated for the same clone, in both orientations, by digestion with exonuclease III and the products of digestion were ligated and then transformed into bacteria. Deletion clones were sequenced by the dideoxy chain termination reaction (Sanger *et al.*, 1977) following the protocol provided with the kit (TaqDye cycle sequencing kit; ABI). Sequencing was performed by using an automated DNA sequencer (model 370 A; ABI). DNA sequence data were assembled and analysed by using the programs DNASIS (Hitachi Software Engineering) and GENETYX (Software Development Co.).

G1, Ic and L viral DNA were cut with *Bam*HI and cloned as described above. Clones were obtained for G1 and Ic only and these were therefore the only two additional variants to be sequenced. Sequencing was done directly with custom-designed primers, 'walking' along the insert in both directions, and it was carried out with an ABI Prism 377 DNA sequencer by using Amplitaq FS and Dye Terminator cycle sequencing ready reaction mix (Perkin Elmer ABI). DNA sequences were compared with each other and with published RTBV sequences. Multiple sequence alignments and preliminary phylogenetic analysis were performed with the Megalign program of the Lasergene software (DNASTAR).

■ **Phylogenetic analysis.** Nucleotide sequences of the six RTBV variants from IRRI and the Serdang isolate were aligned as described above. Phylogenetically informative positions were extracted from the alignment. Visual inspection of the distribution of informative residues in the IRRI variants that were the same as Serdang residues led to the division of the genome into seven segments. One hundred bootstrap samples of each segment were created and analysed by parsimony using programs in the PHYLIP software package version 3.5 (Felsenstein, 1989). PHYLIP programs were also used for maximum-likelihood analyses (Felsenstein, 1981) of the segments and to calculate distances by the Kimura two-parameter method (Kimura, 1980), which were analysed by the Fitch method (Fitch & Margoliash, 1967). Displays of the relationships were created by using TreeView (Page, 1996).

Table 1. Primers used in PCR amplification of specific segments from RTBV variants

Primer locations were based on DNA sequence data for G2 (accession no. AF113831), Phi-1 (Hay *et al.*, 1991) and Phi-3 (Kano *et al.*, 1992). The underlined sequences show an added *Pst*I recognition site. Primers containing this site were used to amplify ORF4 from DNA of all RTBV variants.

Primer	Location	Sequence
P1	6007–6021 (sense)	5′ AGGCTGCAGGAAAAGAGTGCCTAA 3′
P2	6666–6680 (antisense)	5′ CTTCTGCAGAAGATCTTCTCCTTT 3′
P3	6598–6612 (sense)	5′ TATCTGCAGCTGGATACTGGACTG 3′
P4	7268–7282 (antisense)	5′ CATCTGCAGATGTTCCCGCTTTAT 3′
P5	1461–1481 (sense)	5′ GGATATGAACGCCGGTTGTGG 3′
P6	2246–2266 (antisense)	5′ CGGAGACTGATTTATATGCTC 3′
P7	7228–7248 (antisense)	5′ TTGATTCGTACTTAAAGTTGC 3′

Positions at which nucleotide variation occurred were extracted from the alignment. Three kinds of positions at which only two particular residues occurred were identified as potentially useful in examining substitution frequency. Those at which one particular residue occurred in only one isolate were divided into two groups: those in which the unique residue occurred in the Serdang isolate and those in which the unique residue occurred in an IRRI variant. The third group consisted of positions at which both residues occurred more than once and the direction of substitution could be assigned by considering parsimony. The numbers of the twelve possible kinds of substitutions were counted. Substitution frequencies were calculated by summing the events of each substitution class over all isolates and dividing by the mean number of nucleotides of the precursor type in the RTBV genome. The latter was calculated by multiplying the average RTBV base composition by the number of positions in the alignment that were potentially informative for identifying base substitutions. The latter factor corrects for the inability to assign directions for the substitutions in some positions. Extrapolation of RTBV base composition to stability was done with a spreadsheet by repeatedly multiplying base compositions by substitution frequency and calculating new base compositions until the compositions did not change further.

Results and Discussion

Restriction endonuclease mapping and PCR differentiation of RTBV variants

The analysis of restriction endonuclease maps obtained from viral DNAs of G1, Ic, G2 and L variants as well as the Phi-3 variant, used as a control, showed that no endonuclease tested differentiated between variants G1 and Ic. However, *Eco*RV could differentiate between Phi-3, L, G2 and the G1–Ic pair. The actual sizes of fragments were then confirmed from the sequence data for all variants except L (Table 2). Some distinctions could also be made with *Pst*I and *Eco*RI. To check whether RTBV variants could be differentiated on the basis of PCR and RFLP, several regions of the genome were amplified by PCR from viral DNAs of all the five RTBV variants, restricted with *Eco*RI and *Eco*RV endonucleases and checked for strain polymorphism. No differences were found in the restricted PCR products of variants G1 and Ic (data not shown). *Eco*RI digestion of P3/P7 products differentiated variant G2 from the others. *Eco*RI digestion of the P5/P6 products differentiated variants L and G2 from the others, while *Eco*RV digestion of the same products distinguished variants Phi-3 and G2 from the others.

Sequence analysis of G1, Ic and G2

The three variants sequenced differed in length: G1, 8006 bp; Ic, 8005 bp; and G2, 8001 bp. In contrast, the published RTBV sequences of isolates from 1987 (Phi-1) and 1978 (Phi-3) are both 8002 bp (Hay *et al.*, 1991; Kano *et al.*, 1992) and that of 1989 (Phi-2) is 8000 bp (Qu *et al.*, 1991). Alignment of these DNA sequences revealed that the differences in length were due to deletions and insertions in both the coding and the non-coding regions. The most significant difference among the sequences is the insertion of six bases from nt 3825–3830 in both G1 and Ic (Fig. 1). In the non-coding region, Phi-1, Phi-2, Phi-3 and G2 had five adenines from nt 32–36 while G1 had four and Ic had only three. In addition, Phi-1 and Phi-3 had an extra adenine at either nt 7971 or nt 7679, respectively, both of which were absent in the other five sequences.

The base compositions of the positive strands of the seven sequences were not significany different from each other (A = 40%, G = 18%, T = 25% and C = 15%) but were significantly different from those of badnaviruses sugarcane bacilliform virus (EMBL accession no. M89923), Commelina yellow mottle virus (X52938), banana streak virus (AJ002234) and cacao swollen shoot virus (L14546). In particular, RTBV had a higher proportion of adenine and a lower proportion of cytosine compared with the means for these badnaviruses (35% A; 19% C). As with the other pararetroviruses, the base compositions were highly strand-asymmetrical.

The genome organization of the three variants was basically the same as that described previously for RTBV (reviewed by Hull, 1996). As with the sequences of Phi-1, Phi-2 and Phi-3,

Table 2. Positions of the cleavage sites of six different restriction enzymes in seven RTBV variants, as deduced from restriction analysis of viral DNA and confirmed by sequencing

EMBL accession numbers of the sequences included are: AF113830 (G1), AF113832 (Ic), AF113831 (G2), X57924 (Phi-1), M65026 (Phi-2), D10774 (Phi-3) and 076470 (Serdang).

	Restriction sites				
Isolate	*Bam*HI	*Bst*XI	*Eco*RI	*Eco*RV	*Pst*I
G1	7191	4006	122	3440, 4091, 6728	786, 927, 3192, 3893
Ic	7190	4005	121	3439, 4090, 6727	785, 926, 3191, 3892
G2	7186	4001	123, 642, 2134, 7014	1673, 2756, 6723	928, 3888
Phi-1	7186	4001	123, 642, 2134	1673, 2756, 6723	928, 3888
Phi-2	7186	4001	123, 642, 2134, 4867	1673, 2756, 4356, 6723	787, 928, 3888
Phi-3	7186	4001	123, 2134	1673, 2756, 4086, 6723	928, 3888
Serdang	2548, 7186	4007	123	3441, 5402	787, 928, 7186

	911 948
Phi-1	RVTILEHKVEMQNLQDKFETMQIRNKSEITEIP--TTS
Phi-2	RVTILEHKVEMQNLQDKFETMQIRNKSEITEIP--TTS
Phi-3	RVTILEHKVEMQNLQDKFETMQIRNKSEITEIP--TTS
G2	RVTILEHKVEMQNLQDKFETMQIRNKSEITEIP--TTS
G1	RVTILEHK**A**EMQ**K**LQDKFETMQIRNK**P**EITEI**SEA**TTS
Ic	RVTILEHK**A**EMQNLQDKFETMQIRNK**P**EITEI**SEA**TTS
Ser	RVTILEHK**A**EM**HI**LQDKFE**A**MQIRNK**P**EITEI**SEA**TTS

Fig. 1. Alignment of the predicted amino acid residues 911–948 of the seven RTBV isolates showing amino acid changes predicted in the cysteine-rich region of the G1, Ic and Serdang (Ser) isolates. The substitutions are underlined and in bold letters. EMBL accession numbers for the sequences compared are: AF113830 (G1), AF113832 (Ic), AF113831 (G2), X57924 (Phi-1), M65026 (Phi-2), D10774 (Phi-3) and 076470 (Serdang).

the genomes of G1, Ic and G2 consisted of four ORFs, the start and end positions of which varied by one or more nucleotides due to insertions or deletions (data not shown). However, in Ic, a T found at nt 25 of ORF1 created a premature TAA termination codon not present in the other sequences. This segment was sequenced eight times, with identical results.

There were numerous other nucleotide substitutions. Relative to Phi-1, chosen as the reference since it had been shown to be infectious, G1, Ic, Phi-2, Phi-3 and G2 differed by 387, 394, 111, 122 and 135 nucleotides, respectively. In all ORFs, the nucleotide sequences of the six IRRI variants were more related to each other than to that of the Malaysian isolate (Table 3). In the intergenic region (nt 7212–7268), a clear difference was observed between the six IRRI variants and the Malaysian isolate. However, no significant variation was observed among the IRRI variants in that region or in the gap (D2) region. Comparison of the nucleotide identities of the four ORFs of the three newly sequenced variants with each other and with Phi-1, Phi-2, Phi-3 and Serdang revealed that G1 and Ic had almost identical sequences (99·14% identical; Table 3). Both sequences shared common nucleotide substitutions that differentiated them from G2, Phi-1, Phi-2 and Phi-3 (data not shown).

Table 3. Percentage nucleotide identities in ORFs 1–4 and the intergenic region of seven RTBV sequences

Ser, Serdang. See Table 2 for accession numbers of sequences compared.

Comparison	Intergenic region	ORF1	ORF2	ORF3	ORF4
Phi-2/Phi-1	98·5	99·3	98·5	98·5	98·5
Phi-2/G1	97·5	96·2	93·4	94·3	95·1
Phi-2/G2	98·2	98·8	98·5	98·3	97·4
Phi-2/Ic	97·0	96·0	93·4	94·3	95·1
Phi-2/Phi-3	98·1	97·2	98·5	98·3	98·7
Phi-2/Ser	85·8	90·5	93·8	91·5	93·0
G1/Phi-1	97·9	96·5	92·2	94·3	95·6
G1/G2	97·9	96·0	92·2	94·3	95·3
G1/Ic	98·7	99·8	98·2	99·5	99·5
G1/Phi-3	97·7	96·7	92·5	94·2	95·4
G1/Ser	85·3	91·5	92·2	91·0	93·1
G2/Phi-1	98·6	99·5	100	98·3	97·2
G2/Ic	97·3	95·8	91·6	94·3	95·3
G2/Phi-3	98·7	97·0	98·8	98·0	96·5
G2/Ser	86·2	90·3	91·9	91·1	92·6
Ic/Phi-1	97·4	96·3	91·6	94·3	95·6
Ic/Phi-3	97·2	96·5	91·9	94·3	95·4
Ic/Ser	84·6	91·3	92·2	91·0	92·9
Phi-3/Phi-1	98·4	97·5	98·8	98·3	97·9
Phi-3/Ser	90·9	90·2	92·2	91·4	93·3
Phi-1/Ser	85·2	90·8	91·9	91·4	93·0

Some substitutions affected coding in the ORFs. The amino acid sequences of the functional domains (movement domain, integrase domain, lysine-rich region, cysteine-rich region, RNA binding site, protease region, reverse transcriptase region, ribonuclease region) located in ORF3 of the isolates were compared. The cysteine-rich region showed the greatest variation. In that region, both G1 and Ic shared five amino acid substitutions that differentiated them from the other variants (Fig. 1). The amino acid sequence identities for all ORFs mirrored the nucleotide sequence identities, although for the Malaysian isolate, the amino acid identities were higher than the corresponding nucleotide identities, indicating a predominance of synonymous over non-synonymous substitutions.

The observation that DNA sequence variation among the IRRI variants was significantly less than that between those variants and the Serdang isolate from Malaysia is not surprising, since all six variants came from the same virus source maintained in the IRRI greenhouse for many years. As demonstrated by Villegas *et al*. (1997), there are greater variations between field samples collected from different parts of the Philippines than between the 27 full-length clones obtained from the IRRI greenhouse isolate. Similarly, on a larger scale, M. Arboleda and O. Azzam (unpublished results) have shown that RTBV field populations are genetically diverse on the basis of *Eco*RV-digestion of total DNA extracted from infected plants and DNA hybridization with a full-length virus probe. Indeed, experimental evidence indicates that IRRI variants can co-exist. Although the RTBV variants G1 and Ic exhibit cross-protection, the two can be present together in the same plant when inoculated simultaneously. The virus acquired and transmitted by the insect vector from mixed infections exhibited symptoms varying from mild to severe in the differential host, rice cv. FK 135 (Cabauatan *et al*., 1995). This result indicates that the few nucleotide differences between the two isolates are enough to alter symptom expression on certain hosts, provided that both variants are transmitted to FK135 using the same strain of RTSV.

Significance of informative nucleotides between RTBV sequences and phylogenetic analysis

Phylogenetic trees generated from each ORF and the intergenic region by DNASTAR consistently clustered the isolates into three groups: Serdang alone; G1 and Ic; and Phi-1, Phi-2, Phi-3 and G2 (referred to as the G2 cluster) (data not shown). Since the Serdang isolate differed more from the IRRI variants than any variant differed from another, the Serdang sequence was used as a reference to identify the distribution of possibly ancestral nucleotides in each genome. The distribution of Serdang-like residues among phylogenetically informative positions of the variants (Fig. 2) revealed six regions where one of the G2 cluster of four sequences had a higher density of Serdang-like residues than the others. This result suggested that not all regions of the genome had identical evolutionary histories and that recombination had played a role in that evolution. Therefore, the genome was divided into seven segments, segments A and F having anomalous regions in Phi-3, C in Phi-1, D in Phi-2 and G in G2. Segments B and E had no regions in which any isolate had an obviously higher density of Serdang-like residues than the others.

The segments were analysed by bootstrapped parsimony, maximum-likelihood and distance methods. For each segment, the trees generated (Fig. 3) by all three methods were topologically congruent. The trees for all segments were consistent with the clustering observed in the preliminary analysis. However, for segments D and G, the branch separating the G2 cluster from the others had low bootstrap confidence values (66 and 49%, respectively). Within the G2 cluster, the segment trees had different topologies, many of the differentiating branches having significant bootstrap values and branch lengths. As expected from the analysis of Fig. 2, the branches with the variants that had a high density of Serdang-like residues were those closest to the nodes separating the three clusters. The short internal branches of the G2 cluster in the tree for the non-anomalous B segment suggest that a star topology best describes the relationships among the four variants. For the B segment and the non-anomalous segment E, the six variants were, within error, equidistant from the branch to the Serdang isolate, suggesting equal rates of nucleotide substitution for the two clusters since their divergence from a common ancestor. However, the Serdang isolate was 1·4 times as distant from that node in the B tree but 3·9 times as distant in the E tree, suggesting that rates of substitution varied prior to divergence of the variant clusters and that the rates changed differently in the two genome segments. Consistent with the results shown in Table 3, the trees were of different sizes, with that for segment G, containing the intergenic region, being the smallest.

Trees for segments A, C, D, F and G provided modest support for each of the three possible pairwise clusterings of G2, Phi-1, Phi-2 and Phi-3 variants: pairing of G2 and Phi-1 in segment A; Phi-2 and Phi-3 in segments C and G; and G2 and Phi-2 in segments D and F. In all but segment B, one of the four isolates branched significantly more closely (bootstrap values greater than 75%) to the Serdang isolate than the others. The identity of that isolate varied over the length of the genome. Further, the nucleotide sequence identity relationships among the four variants were not consistent over the four ORFs (Table 3). For example, G2 was closest to Phi-1 in ORFs 1 and 2, intermediate in ORF4 and equidistant with Phi-2 in ORF3. These differences suggest an anomalous distribution of the three pairing patterns. Indeed, of 35 positions in which a pair of these variants differed from the other pair, 20 were positions following which the next phylogenetically informative position had the same pattern. Only 13 were expected on the basis of a random distribution of pairing patterns.

The anomalous distribution of Serdang-like residues among phylogenetically informative positions of the variants and the differences in tree topologies for the seven segments suggest that recombination with a sequence not represented among the variants has played a role in the evolution of these variants.

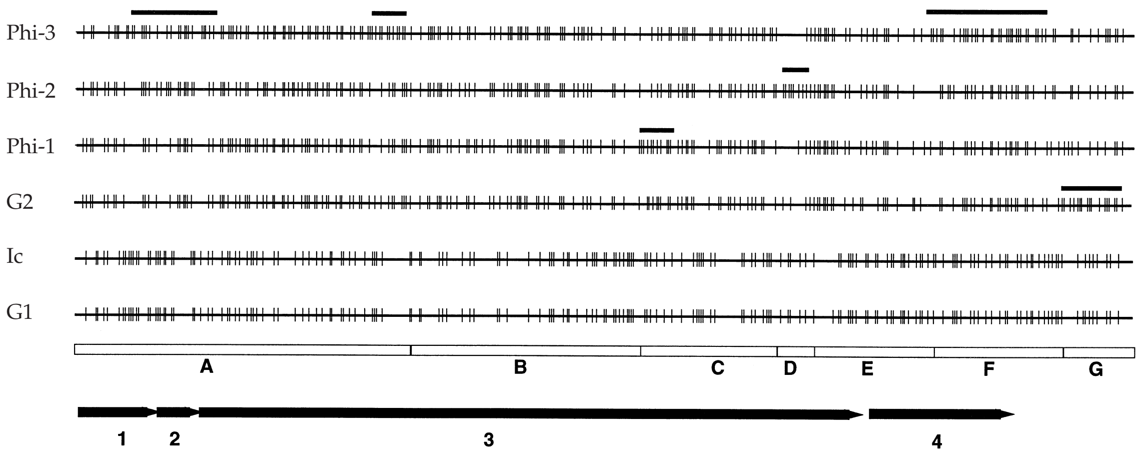

Fig. 2. Distribution of outgroup-specific informative residues among six isolates of RTBV. Vertical lines represent informative positions at which the residue of an isolate was identical to that of the Serdang isolate. Regions of noticeably denser distribution within single isolates are indicated by bars above the distribution. The segmented bar below indicates the division of the aligned genome sequences into segments A–G for phylogenetic analysis. The solid arrows show the positions of RTBV ORFs.

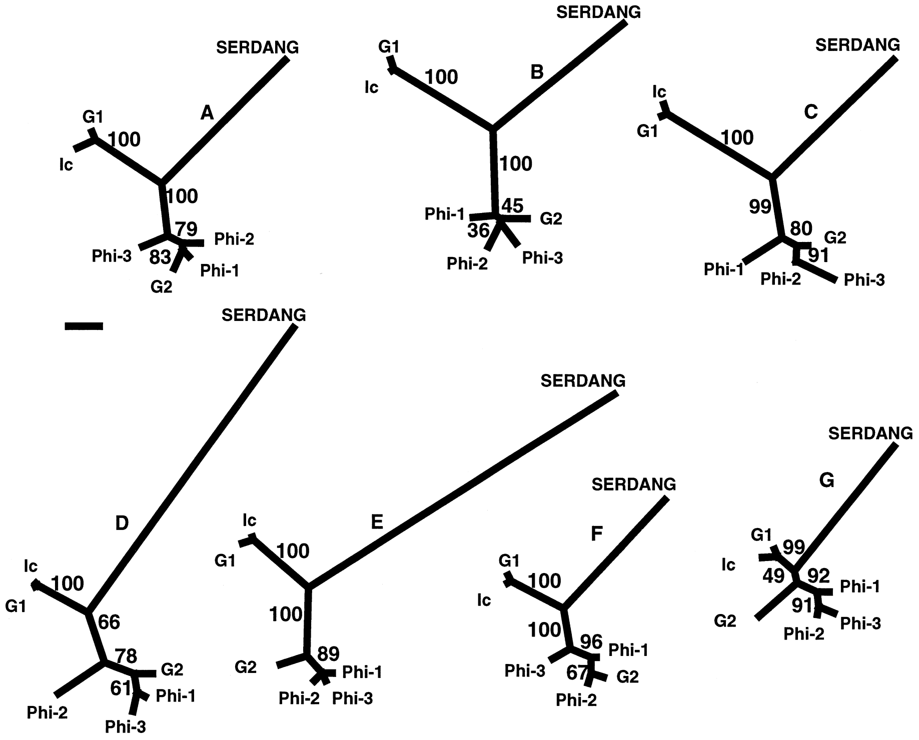

Fig. 3. Maximum-likelihood trees for segments of the RTBV genome. Numbers represent percentages of trees in bootstrapped parsimony analysis that were topologically congruent with the maximum-likelihood trees. All internal branches with bootstrap values were significant ($P < 0{\cdot}01$) by maximum likelihood. Bar represents a distance of 0·01. Sections used (in Phi-1 numbering) were: 1–2500 (A), 2501–3820 (B), 3821–4495 (C), 4496–4833 (D), 4834–5695 (E), 5696–6765 (F) and 6766–8002 (G).

Further examination of the distribution of informative residues suggested that one anomalous stretch was unusually clustered. A set of five contiguous informative positions in isolate Phi-3 was the same as in the Ic–G1 pair. A stretch of this length was expected to occur at random only once in over 380 RTBV genomes. Possibly, the stretch was due to a recombination event between an Ic–G1-like sequence and a Phi-1, Phi-2 and G2-like sequence. The two borders of the recombination event must have been at positions before 626 and after 646. The short sequence possibly exchanged lies near the 3′ end of

ORF1, raising the alternate possibility of functionally determined co-variation at these five positions. All six variants have either one or the other sequence in this region.

Support for recombination events

Averaged over all positions in the genome, the members of the G2 cluster were more diverged from one another than G1 was from Ic. If one assumes that the rates of divergence were equivalent among the six variants, the result implies that G2, Phi-1, Phi-2 and Phi-3 diverged from a common ancestor before G1 and Ic did. Whether this order of divergence is related to the temporal order of their isolation from the IRRI isolate, G1 and Ic having been obtained most recently (Cabauatan *et al*., 1995), remains an open question.

In addition, the anomalous regions, combined with variation in tree topology with map position, suggest that the IRRI greenhouse isolate also contained a third sequence type. The data, then, suggest that G2, Phi-1, Phi-2 and Phi-3 have each undergone recombination with a sequence not represented among the six variants. This recombination must have occurred after their divergence from a common ancestor. Though the anomalous regions were detected by their identity to Serdang isolate residues, the non-represented sequence should not be regarded as a Serdang-like sequence, but merely as one that shares certain ancestral residue types with Serdang. Thus, the IRRI isolate was a mixture of at least three sequence types.

Recombination events have also been implicated in the evolutionary history of cauliflower mosaic virus (CaMV) isolates (Chenault & Melcher, 1994*b*) and observed experimentally after co-inoculation of plants with pairs of marked CaMV DNA sequences (Choe *et al*., 1985; Lebeurier *et al*., 1982; Vaden & Melcher, 1990). These studies suggested that the synthesis of the negative strand by reverse transcription played a major but not exclusive role in generating the recombinants studied. It is likely that some of the putative RTBV recombination events also resulted from strand switching during reverse transcription (Schoelz & Wintermantel, 1993). The recombination event in segment G that gave rise to the G2 variant has ends consistent with it having arisen by negative strand synthesis beginning on the non-represented sequence type and switching templates at the 5′ end of the RNA template to a sequence typical of the G2 cluster. However, no correlations with start, switch or stop sites were obvious for the other putative RTBV recombination events. It should be remembered that, since the recombinations were with a sequence not represented in the collection, the assignment of end-points of the events is very uncertain.

The recombination events implied by the anomalous regions with higher densities of Serdang-like residues may not be the only events that have occurred in the evolution of the IRRI variants. Though no recombination events were detected between G1 or Ic and other sequence types, the data, containing just two variants of this sequence type, cannot rule out such events prior to the divergence of G1 and Ic from a common ancestor or events between G1 and Ic since their divergence. In the G2 cluster, additional recombination events between members probably occurred. Such recombination events could account for the non-random distribution of specific pairs of residue types in the genome. Thus, as long as the variants arising from nucleotide substitutions remain co-resident in the same host, a homogenization of variants may occur, limiting the overall rate of evolution.

Significance of nucleotide substitutions patterns

The numbers of each of the twelve possible kinds of nucleotide substitutions were determined in three ways. When, at a given position, a single IRRI variant differed in its residue type from all the others, a substitution was inferred as being from the predominant type to the unique type. The Serdang-specific unique changes were counted separately, since the root of the tree is probably on the Serdang branch. As a result, the direction of many substitutions would be misinferred, since they occurred on the branch from the root to the non-Serdang isolates. The final method involved counting positions at which more than one variant in a cluster had the same anomalous residue, and parsimony allowed the assignment of substitution direction. The percentages of total substitutions counted in these ways that belong to each class are shown in Table 4. Due to small numbers of events, many differences apparent between the methods of identifying substitutions are not significant. There were, however, several surprising significant differences. In particular, considering only the IRRI variant substitutions, there was a significantly higher percentage of G → A transitions affecting only one variant and significantly higher percentages of T → C transitions and T → G transversions affecting more than one variant. For substitutions affecting only the Serdang isolate, the ratio of transitions to transversions was only 1·9, while it was 4·5 for the IRRI variants. More transversions between A and T accounted for this difference. One interpretation of these differences is that the rates of substitution have not remained the same during divergence of these sequences from a common ancestor.

Since substitutions in just a single IRRI variant were more likely to reflect accurately the probability of one nucleotide changing to another, the possible effect of these substitutions on RTBV DNA base composition was examined. The number of substitutions from A was subtracted from the number of substitutions forming A, and similarly for the other three nucleotides. Only the result of the calculations for T was not significantly different from zero. Gains of G and C and losses of A were predicted. In contrast, when the same calculation was performed on the CaMV data of Chenault & Melcher (1994*a*), no significant changes in the amounts of any nucleotide were predicted.

Given that the analysis of balance in base composition suggested that the base composition was not stable, the

Table 4. Frequencies and percentages of substitution types in RTBV isolates

		Percentage		
Substitution	Frequency*	Single†	Serdang‡	Parsimony§
A → G	9·5 ± 1·7	21·1	18·2	23·7
G → A	18·7 ± 3·7	18·3	11·2	6·9
T → C	17·3 ± 2·9	23·9	18·9	33·1
C → T	21·7 ± 4·3	18·3	17·6	18·6
A → T	2·9 ± 1·0	6·3	9·2	3·8
A → C	1·6 ± 0·7	3·5	4·9	4·7
G → T	1·4 ± 1·0	1·4	0·5	0·9
G → C	0	0·9	–	0·5
T → A	1·5 ± 0·9	2·1	10·7	3·1
T → G	2·0 ± 1·02	1·6	–	2·3
C → A	1·7 ± 1·2	1·9	4·4	2·2
C → G	0·8 ± 0·8	0·7	1·5	0·3
Total number	–	142	391	317

* Number of substitutions of the types listed divided by the mean number of the substituted bases in the RTBV genome, corrected for positions where substitutions could not be assigned. Numbers of substitutions were those calculated from unique nucleotides appearing in IRRI variants.

† Substitutions calculated from unique nucleotides appearing in IRRI variants.

‡ Substitutions calculated from unique nucleotides appearing in the Serdang isolate. Substitutions were characterized as being from IRRI variants to the Serdang isolate, though no direction of substitution can be implied.

§ Substitutions calculated from positions at which each of two residue types occurred more than once and the direction of substitution could be assigned by considering parsimony.

substitution frequencies were used to predict the base composition that would be achieved after repeated replication and mutation, according to the calculated substitution frequencies. The resulting base composition decreased in the fraction of A and increased in the proportion of C. The extrapolated composition was not significantly different from those of other badnaviruses and of CaMV.

The same data were used to calculate substitution frequencies (Chenault & Melcher, 1994*a*), which measure the probability of one nucleotide being replaced by another. The calculation corrects for the base composition of the DNA. The frequencies are shown in Table 4. The frequencies of G → A, T → C and C → T transitions were not significantly different from one another. However, the frequency of A → G transitions was significantly less, a result also obtained for CaMV (Chenault & Melcher, 1994*a*). The calculated RTBV frequencies were similar to the CaMV frequencies, suggesting that approximately the same amount of evolutionary change is represented in the RTBV variants as in the CaMV isolates.

Though the substitutions observed have not significantly affected RTBV base composition, since the accepted changes occurred in less than 5% of positions, base composition changes may have occurred in the distant past and may continue into the future. The base composition of the RTBV positive strand is anomalous among the badnaviruses. The composition is also not that expected from the spectrum of substitutions that have occurred after the divergence of Phi-1, Phi-2, Phi-3 and G2 from one another. The spectrum has been deduced from changes at only a few positions in the RTBV genome. Assuming that the spectrum is applicable to all positions leads to the prediction that, if RTBV is allowed to continue to evolve under the present conditions, its base composition will eventually mirror those of other badnaviruses. Its current composition is unstable in this view.

Implications of these evolutionary forces on the evolution of the IRRI RTBV isolate

The analysis in this study of six RTBV sequences that were derived from the same greenhouse virus culture at IRRI, which was in turn originally obtained from the field, confirms that the IRRI isolate is a mixture of variants with different biological and genetic properties (quasispecies). The six sequences fell into two clusters of sequence types, one consisting of G1 and Ic and the other consisting of the remaining four variants, Phi-1, Phi-2, Phi-3 and G2. Examining genome sequences in a heterogeneous virus population allows us to check the causes and effects of adaptive or evolutionary changes, and the information would be important in developing transgenic plants with resistance to a wide array of virus isolates (Hull, 1994). Alignment of G1, G2 and Ic sequences with those of the other isolates revealed highly conserved regions, particularly in the functional domains in ORF3, and suggested that the cysteine-rich region may be the only domain that carries some adaptive changes. It would be interesting to monitor field populations for that specific domain and examine its relationship to the different adaptive changes seen.

The authors would like to acknowledge that part of this work was supported by the Oklahoma Agricultural Experiment Station.

References

Cabauatan, P. Q. & Hibino, H. (1988). Isolation, purification and serology of rice tungro bacilliform and spherical viruses. *Plant Disease* **72**, 526–528.

Cabauatan, P. Q., Cabunagan, R. C. & Koganezawa, H. (1995). Biological variants of rice tungro viruses in the Philippines. *Phytopathology* **85**, 77–81.

Cabauatan, P. Q., Arboleda, M. & Azzam, O. (1998). Differentiation of rice tungro bacilliform virus strains by restriction analysis and DNA hybridization. *Journal of Virological Methods* **76**, 121–126.

Chenault, K. D. & Melcher, U. (1994*a*). Patterns of nucleotide sequence variation among cauliflower mosaic virus isolates. *Biochimie* **76**, 3–8.

Chenault, K. D. & Melcher, U. (1994*b*). Phylogenetic relationships reveal recombination among isolates of cauliflower mosaic virus. *Journal of Molecular Evolution* **39**, 496–505.

Choe, I. S., Melcher, U., Richards, K., Lebeurier, G. & Essenberg, R. C. (1985). Recombination between mutant cauliflower mosaic virus DNAs. *Plant Molecular Biology* **5**, 281–289.

Fan, Z., Dahal, G., Dasgupta, I., Hay, J. & Hull, R. (1996). Variation in the genome of rice tungro bacilliform virus: molecular characterization of six isolates. *Journal of General Virology* **77**, 847–854.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Felsenstein, J. (1989). PHYLIP – phylogeny inference package. *Cladistics* **5**, 164–166.

Fitch, W. M. & Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155**, 279–284.

Hay, J. M., Jones, M. C., Blakebrough, M. L., Dasgupta, I., Davies, J. W. & Hull, R. (1991). An analysis of the sequence of an infectious clone of rice tungro bacilliform virus, a plant pararetrovirus. *Nucleic Acids Research* **19**, 2615–2621.

Hibino, H., Roechan, M. & Sudarisman, S. (1978). Association of two types of virus particles with penyakit habang (tungro disease) of rice in Indonesia. *Phytopathology* **68**, 1412–1416.

Hibino, H., Ishikawa, K., Omura, T., Cabauatan, P. Q. & Koganezawa, H. (1991). Characterization of rice tungro bacilliform and rice tungro spherical viruses. *Phytopathology* **81**, 1130–1132.

Hull, R. (1994). Resistance to plant viruses: obtaining genes by non-conventional approaches. *Euphytica* **75**, 195–205.

Hull, R. (1996). Molecular biology of rice tungro viruses. *Annual Review of Phytopathology* **34**, 275–297.

Kano, H., Koizumi, M., Noda, H., Hibino, H., Ishikawa, K., Omura, T., Cabauatan, P. Q. & Koganezawa, H. (1992). Nucleotide sequence of capsid protein gene of rice tungro bacilliform virus. *Archives of Virology* **124**, 157–163.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120.

Lebeurier, M., Hirth, L., Hohn, B. & Hohn, T. (1982). In vivo recombination of cauliflower mosaic virus DNA. *Proceedings of the National Academy of Sciences, USA* **79**, 2932–2936.

Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**, 357–358.

Qu, R. D., Bhattacharyya, M., Laco, G. S., De Kochko, A., Suba Rao, B. L., Kaniewska, M. B., Elmer, J. S., Rochester, D. E., Smith, C. E. & Beachy, R. N. (1991). Characterization of the genome of rice tungro bacilliform virus: comparison with Commelina yellow mottle virus and caulimoviruses. *Virology* **185**, 354–364.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, USA* **74**, 5463–5467.

Schoelz, J. E. & Wintermantel, W. M. (1993). Expansion of viral host range through complementation and recombination in transgenic plants. *Plant Cell* **5**, 1669–1679.

Takahashi, Y., Tiongco, E. R., Cabauatan, P. Q., Koganezawa, H., Hibino, H. & Omura, T. (1993). Detection of rice of rice tungro bacilliform virus by polymerase chain reaction for assessing mild infection of plants and viruliferous leafhoppers. *Phytopathology* **83**, 655–659.

Vaden, V. R. & Melcher, U. (1990). Recombination sites in cauliflower mosaic virus DNAs: implications for mechanisms of recombination. *Virology* **177**, 717–726.

Villegas, L. C., Druka, A., Bajet, N. B. & Hull, R. (1997). Genetic variation of rice tungro bacilliform virus in the Philippines. *Virus Genes* **15**, 195–201.

Received 25 January 1999; Accepted 7 April 1999