

Molecular evolution of *Turnip mosaic virus*: evidence of host adaptation, genetic recombination and geographical spread

Kazusato Ohshima,¹ Yuka Yamaguchi,¹ Ryo Hirota,¹ Tamaki Hamamoto,¹ Kenta Tomimura,¹ Zhongyang Tan,¹ Teruo Sano,² Fumio Azuhata,³ John A. Walsh,⁴ John Fletcher,⁵ Jishuang Chen,⁶ Abed Gera⁷ and Adrian Gibbs⁸

¹Laboratory of Plant Virology, Faculty of Agriculture, Saga University, Saga 840-8502, Japan

²Laboratory of Plant Pathology, Faculty of Agriculture & Life Sciences, Hirosaki University, Hirosaki 036-8561, Japan

³Tohoku Seed Co. Ltd, Utsunomiya 321-3232, Japan

⁴Plant Pathology & Microbiology Department, Horticulture Research International, Wellesbourne, Warwick CV35 9EF, UK

⁵Crop & Food Research, Private Bag 4704, Christchurch, New Zealand

⁶Faculty of Life Science, Zhenjiang University, Hangzhou 310029, PR China

⁷Department of Virology, Agriculture Research Organization, The Volcani Centre, Bet Dagan 50250, Israel

⁸Faculty of Science, Australian National University, Canberra, ACT 2601, Australia

Turnip mosaic virus (TuMV), a species of the genus *Potyvirus*, occurs worldwide. Seventy-six isolates of TuMV were collected from around the world, mostly from *Brassica* and *Raphanus* crops, but also from several non-brassica species. Host tests grouped the isolates into one or other of two pathotypes; *Brassica* (B) and *Brassica–Raphanus* (BR). The nucleotide sequences of the first protein (P1) and coat protein (CP) genes of the isolates were determined. One-tenth of the isolates were found to have anomalous and variable phylogenetic relationships as a result of recombination. The 5′-terminal 300 nt of the P1 gene of many isolates was also variable and phylogenetically anomalous, whereas the 380 nt 3′ terminus of the CP gene was mostly conserved. Trees calculated from the remaining informative parts of the two genes of the non-recombinant sequences by neighbour-joining, maximum-likelihood and maximum-parsimony methods were closely similar, and so these parts of the sequences were concatenated and trees calculated from the resulting 1150 nt. The isolates fell into four consistent groups; only the relationships of these groups with one another and with the outgroup differed. The ‘basal-B’ cluster of eight B-pathotype isolates was most variable, was not monophyletic, and came from both brassicas and non-brassicas from southwest and central Eurasia. Closest to it, and forming a monophyletic subgroup of it in most trees, and similarly variable, was the ‘basal-BR’ group of eight BR pathotype Eurasian isolates. The third and least variable group, the ‘Asian-BR’ group, was of 22 BR-pathotype isolates, all from brassicas, mostly *Raphanus*, and all from east Asia mostly Japan. The fourth group of 36 isolates, the ‘world-B’ group, was from all continents, most were isolated from brassicas and most were of the B-pathotype. The simplest of several possible interpretations of the trees is that TuMV originated, like its brassica hosts, in Europe and spread to the other parts of the world, and that the BR pathotype has recently evolved in east Asia.

Author for correspondence: Kazusato Ohshima.

Fax +81 952 28 8709. e-mail ohshimak@cc.saga-u.ac.jp

The nucleotide sequence data have been submitted to the GenBank, EMBL and DDBJ databases and assigned accession numbers AB076407 to AB076556.

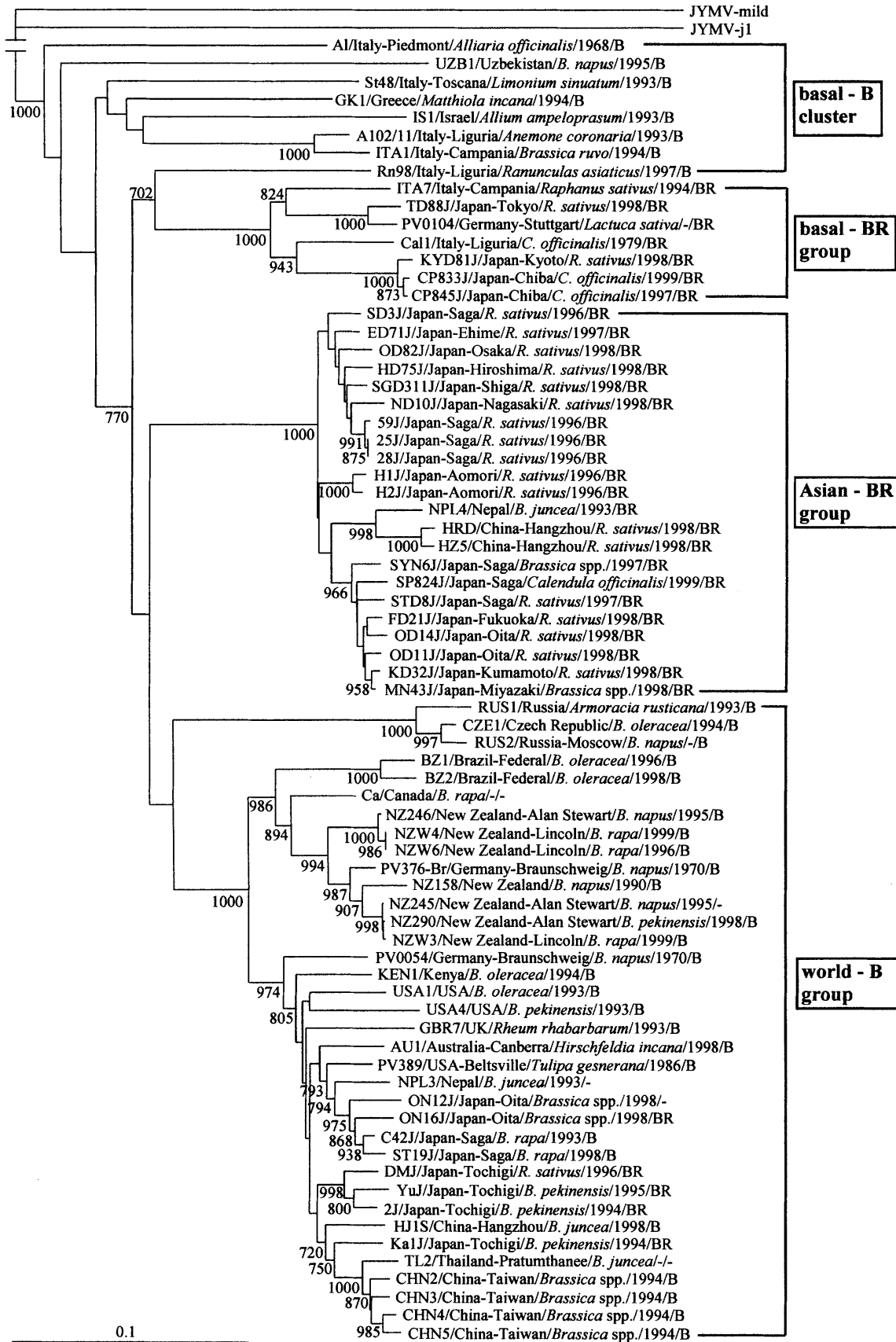


Fig. 1. For legend see facing page.

Introduction

The molecular evolutionary history of viruses is studied not only to satisfy curiosity, but also because such information is essential for the knowledge-based design of strategies for controlling viruses. There have been many such studies of animal viruses seeking evidence of virus origins, mutation rates, selection and fitness, the nature and biological significance of variation, and the mechanisms of reassortment and recombination (Robertson *et al.*, 1995; McCutchan *et al.*, 1996; Sevilla & Domingo, 1996; Holland & Domingo, 1998; Gao *et al.*, 1999; Reid *et al.*, 1999; Gibbs *et al.*, 2001), but fewer and smaller studies of plant viruses (Simon & Bujarski, 1994; Revers *et al.*, 1996; Fraile *et al.*, 1995, 1997; Roossinck, 1997; Gal-On *et al.*, 1998; Masuta *et al.*, 1998; Schneider & Roossinck, 2000).

The nucleotide (nt) sequences of the genes of many different species of plant viruses have been determined, and their phylogenetic relationships have been inferred (Van der Vlugt *et al.*, 1993; Bateson *et al.*, 1994; Tordo *et al.*, 1995; Aleman-Verdaguer *et al.*, 1997; Revers *et al.*, 1997; Roossinck *et al.*, 1999; Bousalem *et al.*, 2000). In some of the studies several isolates of a single species have been studied, and these have shown strain differences that correlate with geographical origins or with strain specialization. We have used this approach in order to understand whether gene sequence analysis can reveal more of the factors that influence worldwide variation within a single virus species, *Turnip mosaic virus* (TuMV).

TuMV has an RNA genome and infects a wide range of plant species, mostly, but not exclusively, from the family Brassicaceae. It is probably the most widespread and important virus infecting both crop and ornamental species of this family, and occurs in many parts of the world including the temperate and tropical regions of Africa, Asia, Europe, Oceania and North/South America (Provvidenti, 1996). TuMV belongs to the genus *Potyvirus*. This is the largest genus of the largest family of plant viruses, the *Potyviridae* (Ward *et al.*, 1995), which itself belongs to the picorna-like supergroup of viruses of animals and plants. TuMV, like other potyviruses, is transmitted by aphids in the non-persistent manner (Shukla *et al.*, 1994). All potyviruses have flexuous filamentous particles 700–750 nm long, each of which contains a single copy of the genome, which is a single-stranded positive sense RNA molecule about 10 000 nt long. The genomes of potyviruses have a single open reading frame that is translated into a single large polyprotein, which is hydrolysed, after translation, into several proteins by virus-encoded proteinases (Riechmann *et*

al., 1992). The genomes of the Canadian (Ca) (Nicolas & Laliberté, 1992) and Japanese (IJ) (Ohshima *et al.*, 1996) isolates of TuMV are 9830 and 9833 nt in length and have single open reading frames which encode polyproteins of 3163 and 3164 amino acids, respectively. Of all potyviral genes, that encoding the coat protein (CP) and situated at the 3'-end of the genome has been most frequently studied for its genetic diversity, whereas there have been fewer studies on the diversity of the first protein (P1) gene, which is the most variable potyviral gene and situated at the 5'-end of the genome (Shukla *et al.*, 1994).

In the study reported here, 76 isolates of TuMV were collected from naturally infected plants throughout the world, and their P1 and CP genes sequenced. Comparisons of these sequences show correlations between their genetic variation and geographical origins, show that some lineages are adapted to particular crop species, and that recombination is a significant generator of the genetic diversity in populations of this virus.

Methods

■ **TuMV isolates.** Seventy-six isolates of TuMV were collected throughout the world by ourselves and colleagues. Details of the isolates, their names, country of origin, original host plant, year of isolation and pathotype are shown in Fig. 1, together with details of two previously sequenced isolates, Canadian (Ca) (Nicolas & Laliberté, 1992) and IJ (Ohshima *et al.*, 1996), that were included in the sequence analyses. In addition, nucleotide sequences of CP genes of two isolates, GK1 and CZE1 (Lehmann *et al.*, 1997), were also included in the analyses. Those from *Brassica* spp. were from Africa, Asia, Europe, Oceania and the Americas, whereas those from *Raphanus* spp. were mainly from Asian countries. Isolates obtained from Asteraceae were collected in Japan, Germany and Italy, whereas the isolates from *Anemone* spp., *Ranunculus* spp. and *Allium* spp. were collected only in Mediterranean countries. All the isolates were inoculated to *Chenopodium quinoa* and serially cloned through single lesions at least three times. They were propagated in *Brassica rapa* cv. Hakatasuwari or *N. benthamiana* plants.

■ **Host tests.** *B. rapa* cv. Hakatasuwari or *N. benthamiana* plants infected systemically with each of the TuMV isolates were homogenized in 0.01 M potassium phosphate buffer (pH 7.0) and mechanically inoculated to young plants of *B. rapa* cv. Hakatasuwari, *B. pekinensis* cvs Nozaki-Igo and Kekkyu Kyoto-3go, *B. napus* cv. Norin-32go, *Raphanus sativus* cvs Taibyo-sobutori and Akimasari. Inoculated plants were kept for at least 4 weeks in a glasshouse at 25 °C. Plants not showing any detectable symptoms were assayed by double antibody sandwich enzyme-linked immunosorbent assay (DAS-ELISA) (Clark & Adams, 1977) using antiserum to isolate 59J.

■ **Cloning of P1 and CP genes.** Viral RNAs were extracted directly from TuMV-infected *B. rapa* or *N. benthamiana* leaves. Complementary DNAs (cDNAs) of the P1 and CP genes of several isolates could not be

Fig. 1. Phylogenetic tree calculated for nt 300–1450 of the concatenated P1 and CP genes of 73 isolates of TuMV (excluding five clear recombinant sequences listed in Table 2 and discussed below; 1J/Japan-Saga/*R. sativus*/1977/BR, CHN1/China-Taiwan/*Brassica* spp./1994/BR, FD271/Japan-Fukuoka/*R. sativus*/1998/BR, HOD517J/Japan-Hokkaido/*R. sativus*/1998/BR and NDJ/Japan-Nagasaki/*R. sativus*/1997/BR). Numbers at each node indicate the percentage of bootstrap samples (1000 replications) (only values > 700 are shown). Horizontal branch lengths are drawn to scale with the bar indicating 0.1 nt replacements per site. The homologous sequences of two isolates (mild and j1) of *Japanese yam mosaic virus* were used as the outgroup. The name of each isolate, its country of origin, original host plant, year of isolation and pathotype are listed.

amplified by RT-PCR, and others did not attain sufficient concentration in RT-PCR for direct sequencing, so their P1 and CP genes were cloned by two methods. In the first, the viral RNA was reverse transcribed and amplified using the Titan One Tube RT-PCR Kit (Roche Diagnostics). The amplified cDNAs were hydrolysed with restriction enzymes using the sites incorporated by the PCR primers, and then separated by electrophoresis in agarose gels. The bands of interest were excised with a razor blade and purified by using the QIAquick Gel Extraction Kit (Qiagen). The eluted DNAs were ligated into alkaline phosphatase-treated pBluescript II SK(+) plasmid vectors (Stratagene), hydrolysed with the appropriate restriction enzymes and used to transform *Escherichia coli* XL-1 Blue (Stratagene). Alternatively, the viral RNAs which could not be amplified by RT-PCR were cloned using random hexamer primers and the TimeSaver cDNA Synthesis Kit (Amersham Pharmacia Biotech), based on the Gubler & Hoffman (1983) method. Synthesized cDNAs were also cloned into pBluescript II SK(+) plasmid vectors that had been hydrolysed by appropriate restriction enzymes.

■ **DNA sequencing.** Nucleotide sequences of the P1 and CP genes of each isolate were determined using three to six cDNA clones. Each cDNA clone was sequenced by 'primer walking' in both directions using the Dye Terminator Cycle Sequencing FS Ready Kit (Applied Biosystems) and an Applied Biosystems DNA Sequencer model 373A. At most we found 2 nt differing between the clones of any one isolate and, when any difference was found, we then sequenced its RT-PCR products directly to determine which was most common. Sequence data were assembled using DNASIS version 3.5 computer program (Hitachi).

■ **Phylogenetic analyses.** The nucleotide sequences were aligned using CLUSTAL W or X (Thompson *et al.*, 1994; Jeanmougin *et al.*, 1998) with default parameters. Their phylogenetic relationships were determined by several methods using the neighbour-joining (NJ) and maximum-likelihood (ML) algorithms of PHYLIP (Version 3.5; Felsenstein, 1993), the maximum parsimony (P) algorithm of PAUP 4.0 beta Version 8 (Swofford, 1998) and the TREE-PUZZLE (Strimmer & von Haeseler, 1996; Strimmer *et al.*, 1997) packages. For NJ analyses, distance matrices were calculated by DNADIST with the Kimura two-parameter option (Kimura, 1980), and trees constructed from these matrices by the NJ method (Saito & Nei, 1987). The homologous regions of the genome of two isolates (mild and j1) of *Japanese yam mosaic virus* (JYMV) (Fuji & Nakamae, 1999) were used as the outgroup for these analyses, as BLAST searches had shown them to be the sequences in the international sequence databases most closely related to those of TuMV. The calculated trees were displayed by TREEVIEW (Page, 1996). The sequences and sub-sequences of them were checked for incongruent relationships that might have resulted from recombination using SISCAN Version 2 (Gibbs *et al.*, 2000; <http://www.anu.edu.au/BoZo/software/>), PHYLPRO (Weiller, 1998) and DIPLOMO (Weiller & Gibbs, 1995). The distance relationships of various sets of sequences were also compared by the DIPLOMO method for evidence of time-dependent and lineage-dependent sequence differences.

For some analyses the aligned P1 and CP genes of each isolate were joined to form a P1/CP sequence 1940 nt long. Then the corresponding parts of two JYMV genomes were aligned with the TuMV sequences by inserting appropriate gaps, and, finally, every position in the aligned sequences with a gap in any sequence was removed. This produced aligned sequences that were 1830 nt long with the P1 sequence represented by the 5'-terminal 969 nt (originally 1086 nt) and the CP sequence by the 3'-terminal 861 nt (originally 864 nt). Trees were calculated by NJ, ML and P methods from nt 300–1450 of these sequences. A bootstrap value for each internal node of the NJ trees was calculated using 1000 random resamplings with SEQBOOT (Felsenstein, 1985) and these values are shown in Fig. 1.

Results

Host specificities

The TuMV isolates were mostly obtained from crops of *Brassica* and *Raphanus* species, but a significant number were from non-brassicaceae, especially Asteraceae. We therefore tested whether there were 'pathotype' differences between the isolates by assessing their ability to infect seedlings of five brassicas; *B. pekinensis* (two cultivars), *B. napus* (one cultivar), *B. rapa* (one cultivar) and *R. sativus* (two cultivars). The results of these tests showed that, although all isolates infected one or more of the four *Brassica* cultivars systemically, they differed consistently in their ability to infect the two *R. sativus* cultivars; those we called the BR (*Brassica*–*Raphanus*) pathotype infected both *R. sativus* cultivars systemically whereas those that did not infect the *R. sativus* cultivars, we called the B (*Brassica*) pathotype. No isolate infected only *R. sativus*. These pathotype categories correlated significantly (Table 1) with the host species from which the isolates were obtained (Fig. 1). All 26 isolates obtained from *R. sativus* samples were the BR pathotype, most (24/32) of those obtained from samples of *Brassica* spp. were of the B pathotype, and 10 of the 15 isolates obtained from non-brassica species were of the B pathotype. Thus *R. sativus* seems to be susceptible only to BR pathotype isolates, whereas *Brassica* spp. and non-brassica species are susceptible to both pathotypes.

The sequences of the different pathotypes were grouped separately, but no consistent difference was found between the B and BR pathotype sequences or the amino acid sequences they encoded.

Gene sequences

In all the TuMV isolates sequenced in this study the P1 gene was 1086 nt in length and the CP gene 864 nt in length. Only a CP gene of an isolate (GK1) previously sequenced by Lehmann *et al.* (1997) was 861 nt. The nucleotide sequence identities for P1 and CP genes between all 78 isolates were 67–100% and 84–100%, respectively. The P1 and CP gene sequences of the IJ isolate (Ohshima *et al.*, 1996), together with their encoded amino acid sequences, were used to search the GenBank database using the BLAST program. The TuMV sequences were found to be consistently more similar to the homologous parts of JYMV than those of any other potyvirus, and when NJ trees were calculated from these sequences, the JYMV sequences formed a group that was itself a sister to all TuMV sequences.

Recombinants

The relationships of the aligned CP genes and of the P1 genes of the TuMV isolates were calculated separately using the NJ, ML and P methods. The resulting trees were similar but

Table 1. Numbers of isolates with different natural host/pathotype combinations

The probability that this distribution of isolates among pathotypes and natural hosts is random is $< 0.001\%$ (χ^2 [Yates' correction] = 29.0; df, 2).

Natural host	Pathotype		Total
	<i>Brassica</i>	<i>Brassica-Raphanus</i>	
<i>Brassica</i> spp.	24	8	32
<i>Raphanus sativus</i>	0	26	26
Other species	10	5	15
Total ...	34	39	73

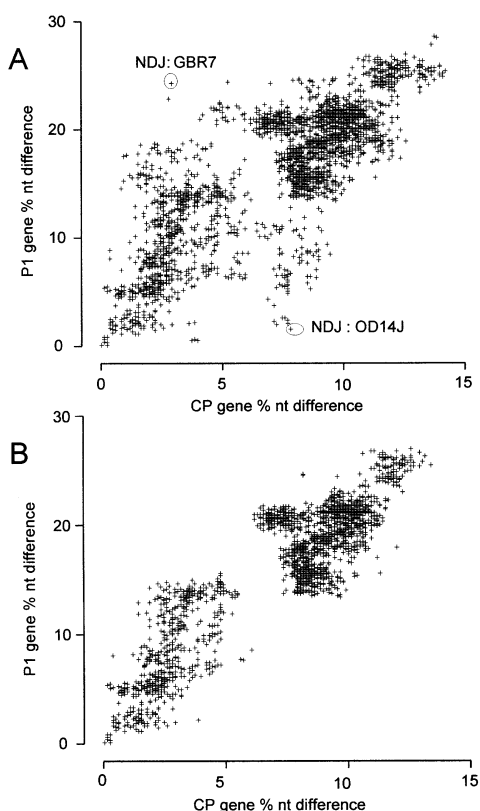


Fig. 2. Graphs showing pairwise comparisons of the patristic distances between P1 sequences plotted against CP comparisons of the same pairs of sequences, either (A) of all 78 sequences or (B) after removing the five clear recombinant sequences listed in Table 2 and discussed below. The points representing particular pairwise comparisons are circled.

showed many inconsistencies in the relative positions of several isolates, and had poor bootstrap support for some lineages (data not shown). These inconsistencies were shown more clearly by plotting the patristic distances between the isolates in the P1 gene tree against their patristic distances in the CP gene tree (Fig. 2A); 'patristic distances' are the distances between taxa within a tree, as distinct from the observed distances between taxa from which the tree was cal-

culated. It can be seen that although the distances between most pairs of P1 genes are twice as much as between their CP genes, there are many pairs that deviate from the mode. For example, in Fig. 2(A) the P1/CP ratio for the NDJ:GBR7 comparison is around 10 but for the NDJ:OD14J comparison less than 1.

These anomalous relationships suggest that some isolates may be recombinants, and so this possibility was investigated in more detail. A preliminary examination of the P1/CP sequences was made by the 'phylogenetic profile' method (Weiller, 1998) using the PHYLPRO program, and clearly showed that parts of some sequences were phylogenetically anomalous (Fig. 3).

Proof that some of the phylogenetic anomalies resulted from recombination rather than convergent selection was obtained by the 'sister scanning' method (Gibbs *et al.*, 2000, 2001). This program allows the separate testing of synonymous, non-synonymous or total nucleotide differences for anomalous relationships. Five sequences, 1J, CHN1, FD27J, HOD517J and NDJ, had statistically significant 'conflicting relatedness signals' (CRSs) in synonymous differences with other sequences, and 1J and FD27J also showed CRSs in their non-synonymous differences. For example, two regions of the 5' part of the 1J sequence (nt 1–100 and 350–680) are significantly (Z -values 3.0) related to the homologous regions of sequence 25J, whereas most of the remaining 3' part of the sequence (nt 690–1350) is significantly related to the homologous region of sequence C42J (Fig. 4A). By contrast, isolate FD27J (Fig. 4B) not only shows significant affinities with sequence 25J, but more closely with sequence ND10J, in the same 5' region and also in all its CP gene, and has a central region (nt 690–980) with affinities to C42J. This indicates that FD27J is, minimally, a double recombinant. Table 2 lists the recombinant parts of the P1/CP genes of each of the five sequences, together with the sequences to which the parts are most closely related.

It is clear that these five recombinant sequences are responsible for many of the pairwise gene comparisons in Fig. 2(A) that are furthest from the main diagonal, as most of these

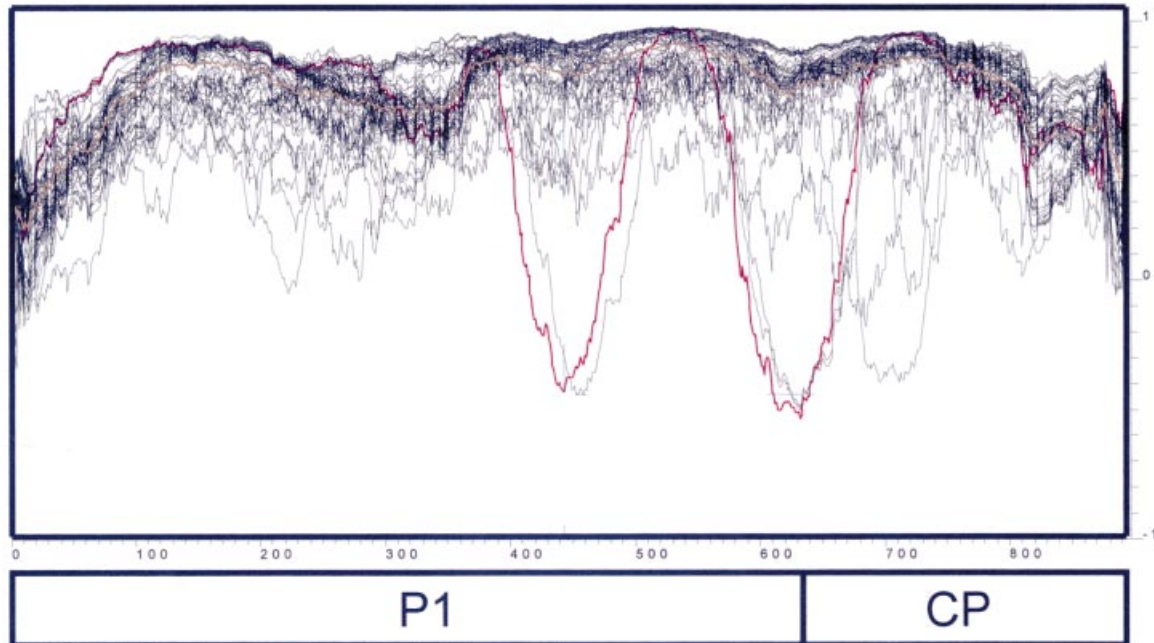


Fig. 3. Graph showing phylogenetic anomalies in 78 TuMV P1/CP sequences detected by the PHYLPRO method; this graph records, for each sequence, the correlations in sequence relationships (% nt differences) with the other 77 sequences in adjacent 100 nt long windows. The P1/CP boundary in the combined sequence of 1830 nt was at nt 861 but, as only 895 of the nt varied, the P1/CP boundary was at nt 615 in the PHYLPRO analysis. Note the obvious anomalies (i.e. correlation minima); those at positions 445 and 615 were for sequence FD27J (red line).

are lost when the five sequences are removed from the comparisons (Fig. 2B) and the linear correlation between P1 and CP patristic distances increases from 0.773, when all 78 P1 and CP sequences are compared, to 0.864 when the five recombinant sequences are omitted.

Many other sequences had phylogenetically inconsistent regions. Some, namely GBR7, HZ5, ITA7, PV0104, Rn98 and St48, had CRSs when examined by SiScan using synonymous differences, but it was not possible to determine which sequence was the recombinant, and which were the sequences closest to the parents, as they had CRSs involving all three pairs of sequences. Others, including GK1, ITA1, NPL4, NZ158, USA4 and UZB1, gave CRSs only when non-synonymous differences were compared, which may indicate that they have sequence regions with convergent similarities resulting from differential selection. When these sequences were omitted from the comparisons of P1 and CP patristic distances, the P1/CP correlation further increased to 0.889.

Comparisons were also made of trees calculated from successive slices of the P1/CP sequences taken along their length and showed that the region between nt 300 and 1450 had the greatest variation (Fig. 5), but was phylogenetically most consistent.

Isolate relationships

We calculated trees from nt 300–1450 of the concatenated P1/CP genes of the 73 isolates, excluding the five clear

recombinants listed in Table 2, using NJ, ML and P methods, and found that they correlated well. For example Fig. 6 shows a DIPLOMO comparison of the NJ (Fig. 1) and ML (data not shown) trees that was obtained by converting each tree to pairwise patristic distances, and then plotting each pairwise distance in the ML tree against that in the NJ tree. It can be seen that the distances between TuMV sequences calculated by the two methods are closely and curvi-linearly related.

In the trees calculated by different methods the isolates fell into four groups. Membership of the groups did not vary and the relationships within each group were closely similar irrespective of the method used. This can be seen, for example, in the comparison between ML and NJ trees in the narrowness of the curved cluster of points in Fig. 6 close to the intersection of the axes of the graph. These within-group clusters were also strongly supported by the bootstrap analyses (Fig. 1). The largest group, 36 isolates mostly from brassicas and of the B-pathotype, was geographically most dispersed so we call it the 'world-B' group. Only five isolates, all of them Japanese, were not B-pathotype. The group was divided into three subgroups. The first subgroup consisted mostly of isolates from the 'New World' (Brazil, New Zealand and Canada). The second subgroup was of isolates from all over the world (Africa, Europe, Australasia and the USA), and third subgroup was of Russian and Czech isolates. All these subgroups were supported by high bootstrap values. The third subgroup was a sister pair to the two other subgroups in the NJ, ML and P trees (data not shown), and we consider this to be strong evidence

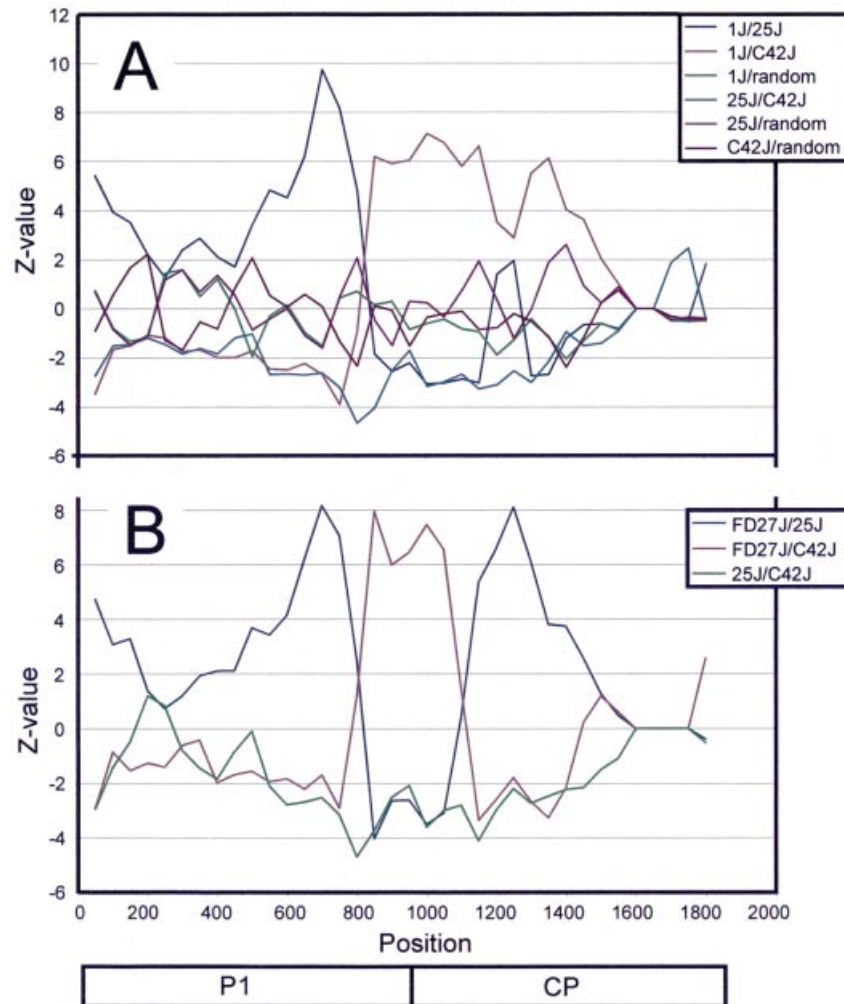


Fig. 4. Graphs showing SiScan analyses. (A) Relationships of the P1/CP sequence of isolate 1J, the two sequences, 25J and C42J, that are closest to its two 'parents', and a sequence constructed of nucleotides chosen at random from the 'parental' sequences during each window comparison. Each window comparison involved sub-sequences of 100 nt, and a step between window positions of 50 nt. Note the strong support (i.e. Z-value 3.0) for 1J being more closely related to 25J than C42J in the 5' 700 nt, and the reverse between nt 700 and 1350. (B) Relationships of the P1/CP sequence of isolate FD27J with the two sequences, 25J and C42J, that are closest to its two 'parents'. Each window comparison involved sub-sequences of 100 nt, and a step between window positions of 50 nt. Note the strong support (i.e. Z-value 3.0) for FD27J being more closely related to 25J than to C42J in the 5' 700 nt and between nt 1000 and 1350, and the reverse between nt 700 and 1000. Comparisons with a sequence constructed by random choice of nucleotides from the 'parental' sequences were omitted for clarity.

that it is part of the world-B group, even though this node was less well supported in the NJ tree by bootstrap sampling.

Next largest was the least diverse group. All 22 isolates came from brassicas, mostly *Raphanus*, all from east Asia mostly Japan, and all are BR-pathotype, so we call this group the 'Asian-BR' group. The other two groupings have the longest branches. One, of eight B-pathotype isolates from both brassicas and non-brassicas from Eurasia, was not monophyletic and, in most trees, was linked directly with the outgroup, so we call it 'basal-B' cluster. Closest to it in most trees, and similarly variable but monophyletic, was the 'basal-BR' group of seven BR pathotype Eurasian isolates. The affinities of isolate Rn98 are uncertain as, although a B-pathotype isolate,

it was placed between the basal-B and basal-BR clusters in NJ, ML and P trees with no clear bootstrap support for inclusion in either of the clusters. It is possible that Rn98 may represent a lineage that included the B-pathotype progenitor of the basal-BR group.

The between-group relationships are more variable than the within-group relationships as shown by the cluster of points in Fig. 6 near the centre of the graph (i.e. comparisons between isolates in different groups) being broader than the cluster closer to the intersection of the axes (i.e. comparisons between isolates in the same group). Nonetheless the 'between-group' cluster shows a clear correlation between the ML and NJ relationships, and the few outgroups mostly

Table 2. Recombinants and 'closest sequence' isolates

Isolate (country of origin)	Pathotype	Natural host	Recombinant type	'Closest sequence' isolates and nt contributed
IJ (Japan)	BR*	<i>Raphanus sativus</i>	BR × B	25J or 59J (Japan) 0–100 & 350–680; C42J (Japan) 690–1350
CHN1 (China)	BR	<i>Brassica</i> spp.	BR × B	HZ5 (China) 0–1100; CHN3 (China) or C42J (Japan) 1150–1400
FD27J (Japan)	BR	<i>Raphanus sativus</i>	BR × B × BR	25J or 59J (Japan) 0–100 & 350–680; C42J (Japan) 690–980; 25J or 59J (Japan) 1000–1350
HOD517J (Japan)	BR	<i>Raphanus sativus</i>	BR × BR	H2J (Japan) 500–980; Ka1J (Japan) 1000–1450
NDJ (Japan)	BR	<i>Raphanus sativus</i>	BR × BR	SYN6J (Japan) 0–100 & 450–980; Ka1J (Japan) 1000–1350

* B, *Brassica*; BR, *Brassica*–*Raphanus*.

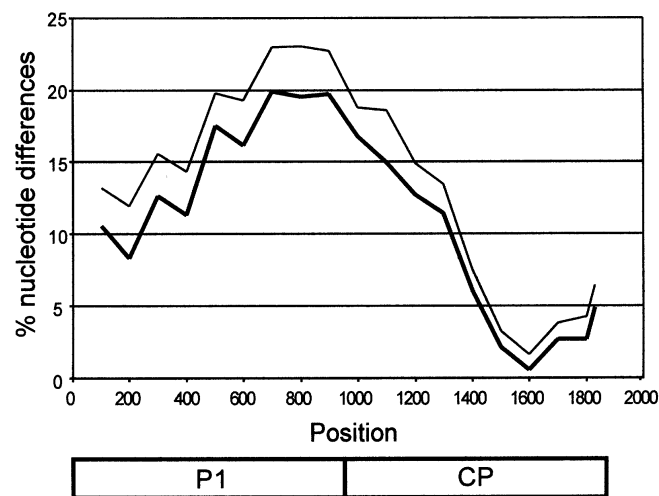


Fig. 5. Graph showing the mean pairwise percentage nucleotide differences for successive 100 nt slices of 78 TuMV P1/CP sequences (bold line) together with those of two *Japanese yam mosaic virus* isolates (narrow line) calculated using the neighbour-joining program.

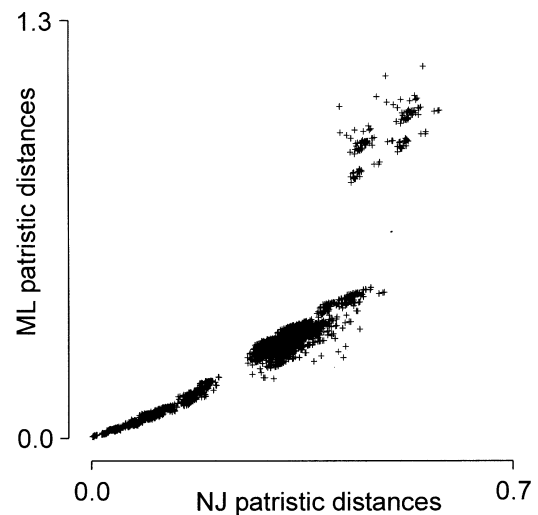


Fig. 6. Graph plotting the distances between all pairs of 73 TuMV and 2 JYMV isolates in a tree calculated by an ML algorithm against the distances in a comparable NJ tree (Fig. 1).

involve isolates with the longest basal branches, such as IS1 and UZB1.

There was a clear geographical pattern within most groups; Chinese, Japanese and New Zealand isolates formed separate groups and these groupings had high bootstrap support (Fig. 1). It was also clear that many of the isolates with 'deepest' branches in the trees were from the Old World, whereas the Asian and Antipodean isolates were in a small number of phylogenetically distinct branches. Similarly, the B pathotype isolates from *Brassica* spp. were phylogenetically more diverse and geographically more dispersed than those from *Raphanus*. The simplest explanation of these patterns is that TuMV probably originated, like its brassica hosts, in

Europe or the Mediterranean region whereas many of the Asian and especially the BR pathotype isolates were from geographically restricted and genetically distinct populations derived from small founding populations that had spread from the centre of origin. A DIPLOMO analysis failed to detect any correlation between the relationships between isolates and the year in which they were isolated (data not shown).

Discussion

This study has shown that genetic recombination is widespread in TuMV populations. The isolates we examined were carefully isolated as single lesions and cloned before analysis to avoid the possibility that the recombination we

detected was an artefact of PCR sequencing, as this has been shown to occur when PCR is used with a mixed target population (Zylstra *et al.*, 1998). To check whether our propagation conditions increased the rate of evolutionary change of TuMV, we passaged 20 isolates 3–24 times and resequenced their P1 and CP genes: there were fewer than two nucleotide differences between the initial and passaged isolates. Therefore, we concluded that the single lesion isolation method that we used did not accelerate changes in the isolate genomes, and hence did not affect our evolutionary analyses.

Specifically designed programs, such as SISCAN, are required to detect the sort of phylogenetic anomalies in sequences that can result from recombination or selection, as the tree-building methods, combined with bootstrapping, that are widely used for phylogenetic analysis may fail to detect recombination (Worobey, 2000). Using such programs we have shown that TuMV populations, like those of an increasing number of other viruses (Bousalem *et al.*, 2000, Padidam *et al.*, 2000, Sanz *et al.*, 2000, Smith *et al.*, 2000), contain recombinants. We found that five of the 78 P1/CP TuMV sequences were clear recombinants (Table 2), and at least another ten may also have been recombinant. However, as the P1/CP sequence represents only about one-fifth of the TuMV genome, the true proportion of recombinants may be much greater, but the true extent of recombination in the world TuMV population will only be known when the full genomic sequences of a representative set of isolates are known.

Our phylogenetic analyses by the NJ, ML and P methods indicate that TuMV, like its principal host family, the Brassicaceae, probably originated and has its major 'centre of diversity' in the Europe–Mediterranean–Asia Minor region. The lineages with the greatest branch lengths (i.e. the basal-B and basal-BR) are probably the oldest, and include many isolates from that region. These groups also include isolates from non-brassica hosts. This may indicate that isolates, optimally adapted to crops of brassicas, spread worldwide in the footsteps of modern agriculture more readily than those adapted to other species, although it could also indicate that the older populations of TuMV are more variable, and hence contain more variants able to infect non-brassica species.

Our phylogenetic analyses also indicate that the original TuMV population was probably B pathotype, but that BR pathotype isolates have evolved from the B pathotype on several occasions; the 'basal-BR' cluster and the 'Asian-BR' group are only of BR pathotype isolates, but there are also BR isolates, all from Japan, in three separate parts of the 'world-B' group. It would be interesting to know whether these solitary BR isolates differ from their nearest B-pathotype relatives in their recombinant status, as this could indicate which part of the genome determines the pathotype.

The fact that *R. sativus* is not susceptible to B pathotype isolates, whereas *Brassica* spp. and non-brassica species are susceptible to both pathotypes suggests that all BR × B recombinants found in *R. sativus* were initially generated in

mixedly infected *Brassica* spp. but were then transmitted to *R. sativus*.

Our analyses of the genetic variation of two genes of the world population of TuMV have revealed its possible region of origin, and indicated the sort of evolutionary changes that may have occurred during its migratory spread, host adaptation and pathogenic segregation, but knowledge of the complete genomic sequences of a very large and representative collection of isolates will be required to confirm whether those conclusions are correct or result from special features of the sample of isolates, and genes, that we studied.

We especially thank Drs R. W. Goldbach (Wageningen Agricultural University, Wageningen, The Netherlands) and A. K. Inoue-Nagata (Virologista, Embrapa-Cenargen, Brasilia, Brazil) for comments on a draft of the manuscript, and for giving encouragement to the project. We also thank many colleagues for supplying TuMV isolates: CP833J and CP845J from Drs S. Uematsu and A. Shioda (Chiba Horticultural Experiment Station, Japan); HOD517J from Dr H. Horita (Hokkaido Ornamental Plants and Vegetable Research Centre, Japan); TL2 from Dr K. Kittipakorn (Department of Agriculture, Bangkok, Thailand); Cal1, Al, A102/11, St48 and Rn98 from Drs V. Lisa and O. Lovisolo (Istituto di Fitovirologia Applicata, Torino, Italy); PV0054 and PV0104 from Dr M. Schöfelder (DSMZ, Braunschweig, Germany); RUS 2 from Dr J. G. Atabekov (Moscow State University, Russia); BZ1 and BZ2 from Drs A. K. Inoue-Nagata and T. Nagata (Embrapa-Cenargen, Brasilia, Brazil); CHN 1–5 from Dr S. K. Green (Asian Vegetable Research and Development Centre, Taiwan); CZE 1, RUS 1 and UZB 1 from Dr J. Špak (Academy of Sciences of the Czech Republic, Ceske Budejovice, Czech Republic); GK 1 from Dr P. Kyriakopoulou (Greece); ITA 1 and ITA 7 from Drs D. Alioto and A. Ragozzino (Universita degli studi di Napoli Federico II, Portici, Italy); and NPL 3-4 from Dr P. Pradanang (Nepal). We especially thank Drs V. Lisa and O. Lovisolo, who sent us plant seeds and several Italian isolates. We wish to thank Drs M. Kusaba (Saga University, Japan) and John Armstrong (Australian National University, Australia) for much help with computer programs and Ms K. Maeda, Y. Ueda, R. Kodama and K. Miki (undergraduate students, Saga University, Japan) for technical assistance. This work was supported by Grant-in-Aid for Scientific Research no. 11660051 from the Ministry of Education, Science and Culture, Japan, and Structure–Function Analysis System of Saga University.

References

- Aleman-Verdaguer, M.-E., Goudou-Urbino, C., Dubern, J., Beachy, R. N. & Fauquet, C. (1997). Analysis of the sequence diversity of the P1, HC, P3, N1b and CP genomic regions of several yam mosaic potyvirus isolates: implications for the intraspecies molecular diversity of potyviruses. *Journal of General Virology* **78**, 1253–1264.
- Bateson, M. F., Henderson, J., Chaleeprom, W., Gibbs, A. J. & Dale, J. L. (1994). Papaya ringspot potyvirus: isolate variability and the origin of PRSV type P (Australia). *Journal of General Virology* **75**, 3547–3553.
- Bousalem, M., Douzery, E. J. P. & Fargette, D. (2000). High genetic diversity, distant phylogenetic relationships and intraspecies recombination events among natural populations of *Yam mosaic virus*: a contribution to understanding potyvirus evolution. *Journal of General Virology* **81**, 243–255.
- Clark, M. F. & Adams, A. N. (1977). Characteristic of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. *Journal of General Virology* **34**, 475–483.

- Felsenstein, J. (1985).** Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.
- Felsenstein, J. (1993).** PHYLIP (Phylogeny Interference Package), version 3.5. Department of Genetics, University of Washington, Seattle, USA.
- Fraile, A., Aranda, M. A. & García-Arenal, F. (1995).** Evolution of the tobamoviruses. In *Molecular Basis of Virus Evolution*, pp. 338–350. Edited by A. Gibbs, C. H. Calisher & F. García-Arenal. Cambridge: Cambridge University Press.
- Fraile, A., Escriu, F., Aranda, M. A., Malpica, J. M., Gibbs, A. J. & García-Arenal, F. (1997).** A century of tobamovirus evolution in an Australian population of *Nicotiana glauca*. *Journal of Virology* **71**, 8316–8320.
- Fuji, S. & Nakamae, H. (1999).** Complete nucleotide sequence of the genomic RNA of a Japanese yam mosaic virus, a new potyvirus in Japan. *Archives of Virology* **144**, 231–240.
- Gal-On, A., Meiri, E., Raccach, B. & Gaba, V. (1998).** Recombination of engineered defective RNA species produces infective potyvirus in planta. *Journal of Virology* **72**, 5268–5270.
- Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M. & Hahn, B. H. (1999).** Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**, 436–441.
- Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000).** Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582.
- Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2001).** Recombination in the hemagglutinin gene of the 1918 ‘Spanish Flu’. *Science* **293**, 1842–1845.
- Gubler, U. & Hoffman, B. J. (1983).** A simple and very efficient method for generating cDNA libraries. *Gene* **25**, 263–269.
- Holland, J. & Domingo, E. (1998).** Origin and evolution of viruses. *Virus Genes* **16**, 13–21.
- Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998).** Multiple sequence alignment with Clustal X. *Trends in Biochemical Sciences* **23**, 403–405.
- Kimura, M. (1980).** A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequence. *Journal of Molecular Evolution* **16**, 111–120.
- Lehmann, P., Petrzik, K., Jenner, C., Greenland, A., Špak, J., Kozubek, E. & Walsh, J. A. (1997).** Nucleotide and amino acid variation in the coat protein coding region of turnip mosaic virus isolates and possible involvement in the interaction with the brassica resistance gene *TuRBO1*. *Physiological and Molecular Plant Pathology* **51**, 195–208.
- McCutchan, F. E., Salminen, M. O., Carr, J. K. & Burke, D. S. (1996).** HIV-1 genetic diversity. *AIDS* **10**, Supplement 3, 13–20.
- Masuta, C., Ueda, S., Suzuki, M. & Uyeda, I. (1998).** Evolution of a quadripartite hybrid virus by interspecific exchange and recombination between replicase components of two related tripartite RNA viruses. *Proceedings of the National Academy of Sciences, USA* **95**, 10487–10492.
- Nicolas, O. & Laliberté, J.-F. (1992).** The complete nucleotide sequence of turnip mosaic potyvirus RNA. *Journal of General Virology* **73**, 2785–2793.
- Ohshima, K., Tanaka, M. & Sako, N. (1996).** The complete nucleotide sequence of turnip mosaic virus RNA Japanese strain. *Archives of Virology* **141**, 1991–1997.
- Padidam, M., Sawyer, S. & Fauquet, C. M. (2000).** Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218–225.
- Page, R. D. (1996).** TreeView: an application to display phylogenetic trees on personal computer. *Computer Applications in the Biosciences* **12**, 357–358.
- Provvidenti, R. (1996).** Turnip mosaic potyvirus. In *Viruses of Plants*, pp. 1340–1343. Edited by A. A. Brunt, K. Crabtree, M. J. Dallwitz, A. J. Gibbs & L. Watson. Wallingford, UK: CAB International.
- Reid, A. H., Fanning, T. G., Hultin, J. V. & Taubenberger, J. K. (1999).** Origin and evolution of the 1918 ‘Spanish’ influenza virus hemagglutinin gene. *Proceedings of the National Academy of Sciences, USA* **96**, 1651–1656.
- Revers, F., Le Gall, O., Candresse, T., Le Romancer, M. & Dunez, J. (1996).** Frequent occurrence of recombinant potyvirus isolates. *Journal of General Virology* **77**, 1953–1965.
- Revers, F., Lot, H., Souche, S., Le Gall, O., Candresse, T. & Dunez, J. (1997).** Biological and molecular variability of lettuce mosaic virus isolates. *Phytopathology* **87**, 397–403.
- Riechmann, J. L., Laín, S. & García, J. A. (1992).** Highlights and prospects of potyvirus molecular biology. *Journal of General Virology* **73**, 1–16.
- Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. (1995).** Recombination in HIV-1. *Nature* **374**, 124–126.
- Roossinck, M. (1997).** Mechanisms of plant virus evolution. *Annual Review of Phytopathology* **35**, 191–209.
- Roossinck, M. J., Lee, Z. & Hellwald, K.-H. (1999).** Rearrangements in the 5′ nontranslated region and phylogenetic analyses of cucumber mosaic virus RNA 3 indicate radial evolution of three subgroups. *Journal of Virology* **73**, 6752–6758.
- Saitou, N. & Nei, M. (1987).** The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425.
- Sanz, A. I., Fraile, A., García-Arenal, F., Zhou, X., Robinson, D. J., Khalid, S., Butt, T. & Harrison, B. D. (2000).** Multiple infection, recombination and genome relationships among begomovirus isolates found in cotton and other plants in Pakistan. *Journal of General Virology* **81**, 1839–1849.
- Schneider, W. L. & Roossinck, M. J. (2000).** Evolutionarily related Sindbis-like plant viruses maintain different levels of population diversity in a common host. *Journal of Virology* **74**, 3130–3134.
- Sevilla, N. & Domingo, E. (1996).** Evolution of a persistent aphthovirus in cytolytic infections: partial reversion of phenotypic traits accompanied by genetic diversification. *Journal of Virology* **70**, 6617–6624.
- Shukla, D. D., Ward, C. W. & Brunt, A. A. (1994).** Introduction. In *The Potyviridae*, pp. 1–26. Edited by D. D. Shukla, C. W. Ward & A. A. Brunt. Wallingford, UK: CAB International.
- Simon, A. E. & Bujarski, J. J. (1994).** RNA–RNA recombination and evolution in virus-infected plants. *Annual Review of Phytopathology* **32**, 337–362.
- Smith, G. R., Borg, Z., Lockhart, B. E., Braithwaite, K. S. & Gibbs, M. J. (2000).** Sugarcane yellow leaf virus: a novel member of the *Luteoviridae* that probably arose by inter-species recombination. *Journal of General Virology* **81**, 1865–1869.
- Strimmer, K. & von Haeseler, A. (1996).** Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* **13**, 964–969.
- Strimmer, K., Goldman, N. & von Haeseler, A. (1997).** Bayesian probabilities and quartet puzzling. *Molecular Biology and Evolution* **14**, 210–211.
- Swofford, D. L. (1998).** PAUP. Phylogenetic analysis using parsimony. Version 4. Sinauer Associates, Sunderland, MA, USA.

- Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994).** CLUSTAL W. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.
- Tordo, V. M.-J., Chachulska, A. M., Fakhfakh, H., Le Romancer, M., Robaglia, C. & Astier-Manificier, S. (1995).** Sequence polymorphism in the 5' NTR and in the P1 coding region of potato virus Y genomic RNA. *Journal of General Virology* **76**, 939–949.
- Van der Vlugt, R. A. A., Leunissen, J. & Goldbach, R. (1993).** Taxonomic relationships between distinct potato virus Y isolates based on detailed comparisons of the viral coat proteins and 3'-nontranslated regions. *Archives of Virology* **131**, 361–375.
- Ward, C. W., Weiller, G. F., Shukla, D. D. & Gibbs, A. (1995).** Molecular systematics of the Potyviridae, the largest plant virus family. In *Molecular Basis of Virus Evolution*, pp. 477–500. Edited by A. Gibbs, C. H. Calisher & F. García-Arenal. Cambridge: Cambridge University Press.
- Weiller, G. F. (1998).** Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Molecular Biology and Evolution* **15**, 326–335.
- Weiller, G. F. & Gibbs, A. (1995).** DIPLOMO: the tool for a new type of evolutionary analysis. *Computer Applications in the Biosciences* **11**, 535–540.
- Worobey, M. (2000).** Extensive homologous recombination among widely divergent TT viruses. *Journal of Virology* **74**, 7666–7670.
- Zylstra, P., Rothenfluh, H. S., Weiller, G. S., Blanden, R. V. & Steele, E. J. (1998).** PCR amplification of murine immunoglobulin germline V genes: strategies for minimization of recombination artefacts. *Immunology and Cell Biology* **76**, 395–405.

Received 7 December 2001; Accepted 11 February 2002