Arch Virol (2005) 150: 1591–1605
DOI 10.1007/s00705-005-0510-4

Archives of Virology
Printed in Austria

Heterogeneity in codon usages of sobemovirus genes

H. Zhou[1,2], **H. Wang**[1], **L. F. Huang**[1], **M. Naylor**[1], and **P. Clifford**[2]

[1]Natural Environmental Research Council/Centre for Ecology and Hydrology Oxford (formerly Institute of Virology and Environmental Microbiology), Oxford, U.K.
[2]Department of Statistics, University of Oxford, Oxford, U.K.

Received August 26, 2004; accepted January 25, 2005
Published online April 18, 2005

Summary. When conventional phylogenetic trees were built using 14 genome sequences of 9 sobemoviruses, two main lineages were apparent: monocot-infecting viruses and dicot-infecting viruses. To investigate whether members of the genus *Sobemovirus* originated from monocot hosts or from dicot hosts, we constructed relationship trees based on Relative Synonymous Codon Usage (RSCU) of the viruses. The RSCU relationship trees grouped the monocot-infecting and dicot-infecting viruses even better than the genome phylogenetic trees. The RSCU approach also enabled direct comparisons among viral and host species. When host species were added into the RSCU tree, the viral species clustered with the monocot hosts, indicating codon usage homologies to monocots. The stability of the RSCU tree was improved when RSCU values were calculated for individual viral open reading frames (ORFs). Most interestingly, the codon usages of the viral ORF-2 that encodes the replicase showed affinity to that of the plants whereas codon usages of the other viral ORFs were not relevant to the host species. All ORF-2s from 3 monocot viruses and 4 out of 6 dicot viruses had greater RSCU affinities to sequences of ORFs in monocot than to dicot hosts, possibly indicating that ORF-2, and therefore the replicase module of sobemovirus has a monocot origin.

Introduction

Phylogenetic analyses based on homologies between sequences have been commonly used in many aspects of bioinformatics. Codon usage bias in viral sequences also provided insights into host adaptation in some cellular pathogens [14]. The codon usage adaptation hypothesis predicts that cellular pathogens, including viruses, will change their codon usage bias towards that of the host, to facilitate utilisation of the host translation machinery with maximal efficiency. Alternatively, observed codon usage variation may also be caused by mutational bias that derived from synonymous mutagenesis.

Mutational bias rather than translational selection has recently been reported to be the more important source of codon usage bias observed in human RNA viruses [5]. Analyses of sequences of both RNA and DNA plant viruses led to a similar conclusion [1]. Such negative findings for translational selection pressure on viruses, offer an opportunity to excavate 'genetic fossils' preserved in viral codon bias and thereby to shed light on the origins of related viruses. Indeed, one of the long-standing hypotheses regarding the origins of viruses, is that viruses and viral genes developed from constituents of normal or degenerate host cells. Different modules of viral genes are also considered to have different origins [4]. With the additional finding that viral codon bias is not affected by gene size in both plant and human RNA viruses [1, 5], these hypotheses can be tested by comparing codon usages between hosts and viruses. Viruses that infect green vascular plants are even more attractive subjects than animal viruses because monocots and dicots have distinguishable codon bias [17].

Many plant virus genera have monocot-infecting as well as dicot-infecting members. Phylogenetic analyses based on sequence homology have progressed tremendously in recent years, revealing the evolutionary relationships within and between viral species. However, very little is known about the origin of any viral genus. Whether a viral genus was originated in a monocot or dicot is an interesting evolutionary question to answer.

The genus *Sobemovirus* is named after the type species, *Southern bean mosaic virus* (SBMV) which has a positive-sense single-stranded RNA (+ss RNA) genome. In general, sobemoviruses have narrow host ranges and lack within-species recombination [3, 8]. A comprehensive review of this genus had been made by Tamm & Truve [16]. In brief, sobemoviruses have a small genome of more than 4 Kb. The genomic RNA links to a genome-linked protein (VPg) at the 5′-end. Flanked by the 5′- and 3′-end untranslated regions (UTR), the viral genes are organized as, ORF-1 encoding P1 that has a gene silencing suppressor function, ORF-2 encoding a polyprotein containing a serine protease at the N-terminus, followed by the VPg then RNA dependent RNA polymerase (RdRp) at the C-terminus, ORF-3 encoding a putative protein which function is unknown, and ORF-4 encoding the viral coat protein (CP). Two types of ORF-2 are found in sobemoviruses: (A) continuous ORF, and (B) -1 reading frame shift for the RdRp. Sobemo-RdRp sequences characteristically represent a so-called Sobemo-lineage in the Supergroup-I of viral RdRp genes [6]. The putative ORF-3 is nested within the A-type ORF-2 and is suspected to be translated by -1 frame shifting. Expression of ORF-4 (CP) is mediated by a subgenomic RNA.

In addition to SBMV [7, 12, 18], there are 3 monocot-infecting and 5 dicot-infecting sobemoviruses that were fully sequenced towards the end of 2003 [3, 8, 9, 11, 13]. *Rice yellow mottle virus* (RYMV) has been compared for codon usage similarity with its natural host, rice [1]. *Turnip rosette virus* (TRoV) can infect *Brassica* species and *Arabidopsis thaliana* (unpublished data from our lab, and from information associated with GenBank:AY177608 submitted by Callaway, Thornesbury, and Lommel 2002). In this study, we used the sequences of ORFs of *Oryza sativa*, and *A. thaliana* from the publicly available TIGR website, and

the sequences of ORFs in sobemoviruses that have been fully sequenced, to draw codon usage trees that revealed a likely monocot origin of sobemovirus ORF-2. We also used the codon usage tables available for 17 plant species in the website of Kazusa DNA Research Institute, to testify the stability of the codon usage trees.

Materials and methods

The three datasets used in this study were collected from publicly available websites before October 2003.

DATASET 1: Comprised the protein-coding nucleotide sequences of the plant genes (intron-less, no untranslated region) were downloaded from the ftp site of TIGR (ftp://ftp.tigr.org), which includes

O. sativa: (112 MB, for chromosomes 1–12, Table 1)

ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/BAC_PAC_clones/OSA1.cds

A. thaliana: (38.5 MB, for chromosomes 1–5, Table 1)

ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/a_thaliana/annotation_dbs/ATH1.cds

DATASET 2: Consisted of 14 complete genome sequences of 9 *Sobemovirus* (Table 2) were downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/). The sequences were directly used for the construction of genome phylogenetic trees. Then each complete viral genome sequence was separated into 3 or 4 ORFs, according to the descriptions associated with the viral genome sequences, for constructing codon usage cluster trees. Three viral sequences (CfMV, TRoV, SCMoV) lacked a recognisable ORF-3.

Table 1. Number of *A. thaliana* and *O. sativa* protein-coding genes used in this study (downloaded from ftp://ftp.tigr.org)

Species	Chromosome	Number of CDS
Arabidopsis thaliana	1	7278
	2	4597
	3	5665
	4	4413
	5	6629
Oryza sativa	1	10478
	2	8983
	3	7535
	4	7886
	5	6751
	6	7321
	7	6748
	8	6413
	9	5225
	10	4861
	11	5247
	12	5670

Table 2. Complete genome sequences of *Sobemovirus* used for this study (downloaded from http://www.ncbi.nlm.nib.gov/)

Sequence ID	Virus name	Host
AB040447	Cocksfoot mottle virus, Japan (CfMV-J)	Monocot
Z48630	CfMV, Finland (CfMV-F)	Monocot
L40905	CfMV, Russia (CfMV-R)	Monocot
U23142	Rice yellow mottle virus, Nigeria (RYMV-N)	Monocot
L20893	RYMV	Monocot
AB040446	Ryegrass mottle virus (RGMoV)	Monocot
AY177608	Turnip rosette virus (TRoV)	Dicot
U31286	Lucerne transient streak virus (LTSV)	Dicot
AF208001	Subterranean clover mottle virus (SCMoV)	Dicot
M23021	Southern cowpea mosaic virus (SCPMV)	Dicot
AY004291	Sesbania mosaic virus (SeMV)	Dicot
L34672	Southern bean mosaic virus, Bean strain (SBMV-B)	Dicot
AF055888	SBMV-ARK-S	Dicot
AF055887	SBMV-ARK-B	Dicot

DATASET 3: Was Codon Usage Tables in 12 dicot species and 5 monocot species (Table 3) downloaded from the website of Kazusa DNA Research Institute (http://www.kazusa.or.jp/) and converted to Relative Synonymous Codon Usage (RSCU) for the construction of codon trees. The selected species were each represented by more than 20,000 codons (Table 3).

Table 3. Codon usage tables used in RSCU trees (downloaded from http://www.kazusa.or.jp/)

	Species	Number of CDS*	Number of codons*
Dicot	*Arabidopsis thaliana*	66749	26545376
	Brassica juncea	103	39594
	Brassica napus	474	175396
	Brassica oleracea	119	41455
	Brassica rapa	139	39948
	Lactuca sativa	37	21150
	Medicago sativa	227	81596
	Medicago truncatula	223	97441
	Nicotiana tabacum	1226	462709
	Solanum tuberosum	591	247205
	Vigna radiata	76	26912
	Vigna unguiculata	66	25227
Monocot	*Avena sativa*	63	27762
	Hordeum vulgare subsp. vulgare	615	232294
	Oryza sativa	48914	18417486
	Triticum aestivum	874	320084
	Zea mays	1863	794495

*Information provided together with the codon usage tables

Viral genome sequences were aligned by using the ClustalX (Version 1.81) program. phylogenetic trees were constructed using the Maximum Parsimony (MP) method (with 1000 bootstraps) provided in MEGA (2.1), and the Maximum Likelihood (ML) method provided in PHYLIP (3.573c) and viewed by TreeView (1.6.6) in phylograms.

Codon usage cluster trees were produced by CodonW (1.4.2) and S-Plus (6.1) programs. Briefly, a codon cluster tree was constructed by four steps. (A) calculating the RSCU values for a group of ORFs or each individual ORF (by CodonW), (B) constructing a raw RSCU matrix for each viral and plant species, by using the RSCU tables of each species using the format for S-Plus (a row was for a species and a column was for a codon), (C) calculating the distance matrix, and (D) producing the codon usage cluster tree according to the distance (by S-Plus).

More than 10 kinds of indices divisible into two classes were commonly used to measure the difference between observed and expected codon usages. One measures the overall deviation of codon usage, for example the Nc, GC, and RSCU indices, and the other measures the bias towards a particular subset of preferred codons, for example the Fop, CAI and CBI indices. RSCU is calculated as the ratio of observed frequency of a codon to the frequency expected if codon usage was uniform within a synonymous codon group [14]. Compared with other indices, such as Nc or GC3s, RSCU is more sensitive to the component of synonymous codon. For synonymous codon i of an n-fold degenerate amino acid, it is estimated as:

$$RSCU_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

Where X_{i} is the number of the occurrences of codon i, and n is 1, 2, 3, 4, 5, or 6. In this study, RSCU values of DATASET 1 and DATASET 2 were generated by CodonW in a formal format. DATASET 3 was downloaded as Codon Usage Tables for each plant species, and their RSCU values were calculated with the above formula.

The RSCU table (Table 4) were formatted in Microsoft Excel as: rows were each for an species (viral or host), and columns were each for a particular codon. Note, TGG and ATG have been omitted in the table, because they do not have synonymous codons, therefore their RSCU values were constant (always 1) and valueless for calculating the distance matrices. The stop codons, UAA, UAG and UGA, were omitted as in other studies [1]. Therefore the RSCU tables contained 59 columns.

The RSCU table was then directly converted to X, a $m \times 59$ raw multivariate data matrix [m species (the viral genes or genomes, and plant chromosomes of interest), and 59 codons]. Here, x_{ij} denotes the entry in the i-th row and j-th column of matrix X and means the RSCU value of j-th Codon in i-th specie. Furthermore, a 59-dimensional space can be constructed by these 59 codons and each species can be looked as a vector in this space.

Table 4. RSCU table

	UUU	UCU	UAU	GUG	...	GCG	GAG	GGG
AB040446	0.42	0.90	0.69	1.38	...	1.17	1.07	1.01
AB040447	0.79	1.02	0.86	1.32	...	0.94	1.12	0.55
AF055887	0.87	1.40	1.00	0.97	...	0.86	1.19	0.75
AF055888	0.84	1.39	0.96	0.97	...	0.83	1.16	0.78
...	...	...	...	...	...	...	...	...

RSCU raw matrix X with m species and 59 codons

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{159} \\ x_{21} & x_{22} & \dots & x_{259} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{m59} \end{bmatrix}$$

Based on the multivariate data matrix X, distance matrix D (a $m \times m$ matrix) was then calculated with the *Euclidean distance* formula by the S-Plus program. Each pairwise distance, such as d_{ij} or d_{ji} represented the distance between two species, i and j, in the particular 59-dimensional space.

Euclidean distance formula in 59 dimensions is given by:

$$d_{ij} = \sqrt{\sum_{k=1}^{59} (x_{ik} - x_{jk})^2}$$

Where x_{ik} and x_{jk} are the RSCU values of the k-th codon in species i and j respectively.

Distance matrix D, $d_{ij} = d_{ji}$, $(i \neq j)$ and $d_{ii} = 0$ $(i = 1, \dots m)$

$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1m} \\ d_{21} & d_{22} & \dots & d_{2m} \\ \dots & \dots & \dots & \dots \\ d_{m1} & d_{m2} & \dots & d_{mm} \end{bmatrix}$$

Group Average was then taken to be the inter-group dissimilarity measure to draw the codon usage agglomerative hierarchical clustering trees, by S-Plus.

Group Average formula is given by:

$$s_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{i \in B} d_{ij}$$

Where s_{AB} is the value of dissimilarity between Group A and B, the average distance between the species of Group A and Group B in this particular 59-dimensional space. n_A and n_B are the numbers of species in Group A and B respectively, and d_{ij} is the entry in the i-th row and j-th column of the above distance matrix D.

The above two steps were performed using the S-Plus program. The Excel RSCU table was imported into S-Plus as a data set (data) containing all rows (except the 1st row) of the 2nd column to the last column, and an index set (index) containing all rows (except the 1st row) of the 1st column. The data set was used to calculate the distance by using a command:

h←hclust(dist(data, metric = "euclidean"), method = "average")

The index set was used to label the RSCU tree by using a second command:

plclust(h, col = 1, labels = index$C1)

Such a tree reveals the codon usage relationships of these species. The greater the similarity of synonymous codon usage between two species, the closer is their *Euclidean distance*, and the closer two branches appear in the tree. A scale is produced by S-Plus on the right side of each figure. This is the *Euclidean distance* and based on the RSCU values. The bioinformatical meaning is obscure but it helps visualize the relative distance between branches.

Results

The MP genome tree (Fig. 1A) shows that two of the three monocot-infecting viral species (labelled with MI), CfMV and RYMV were closer to each other than to the dicot-infecting viruses (labelled with DI) as previously reported [8]. The third monocot-infecting virus, RGMoV was assosiated with the dicot viruses with low

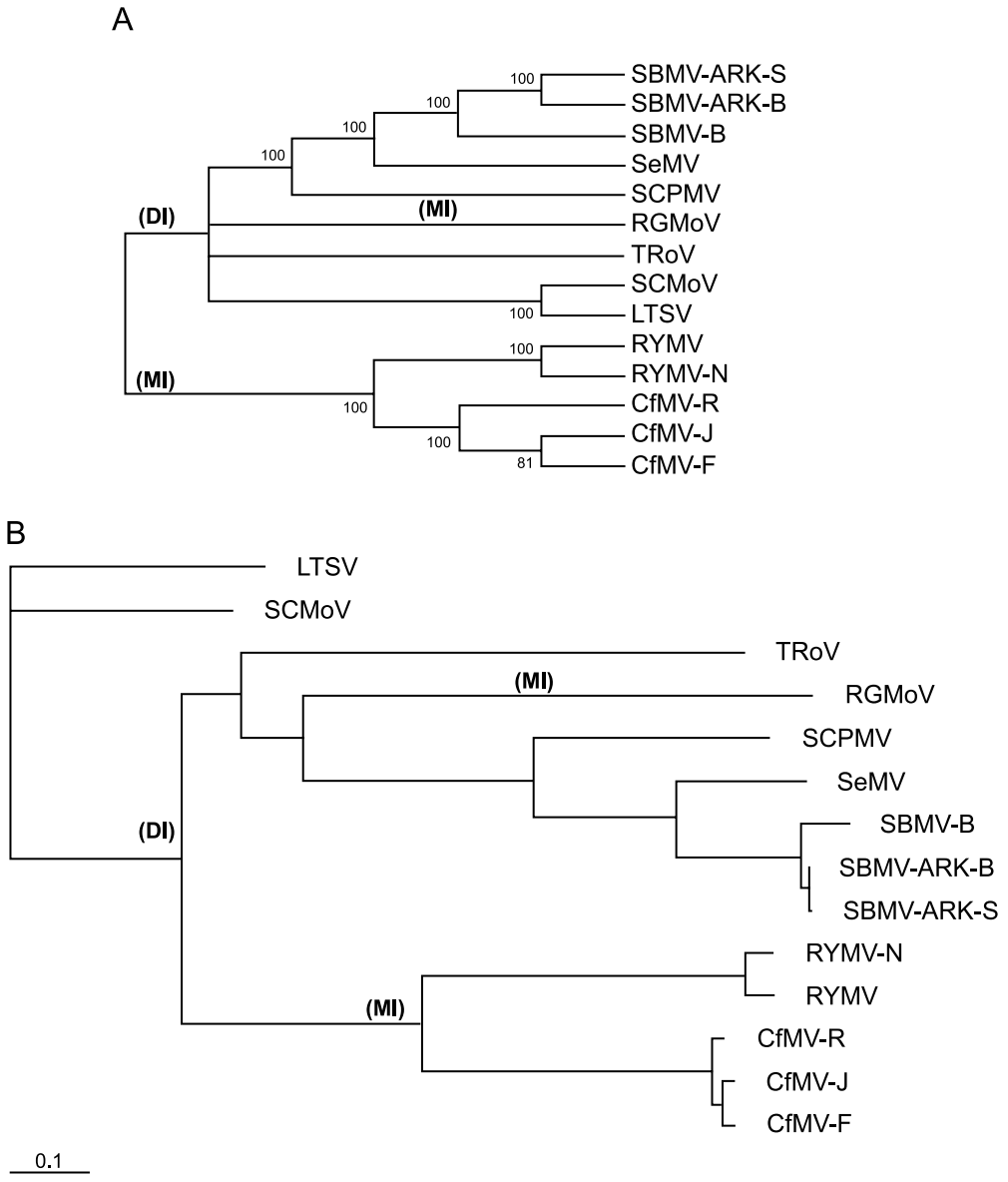

Fig. 1. Phylogenic genome trees of sobemoviruses. Dicot- and monocot-infecting viruses were labelled with DI and MI, respectively. **A**: Bootstrap (1000) consensus tree (cut-off value of 50%) using the Maximum Parsimony (MP) method. The numbers on branches showed the confidence level. **B**: Unrooted genome phylogram using the Maximum Likelihood (ML) method. The bar showed the scale of branch length

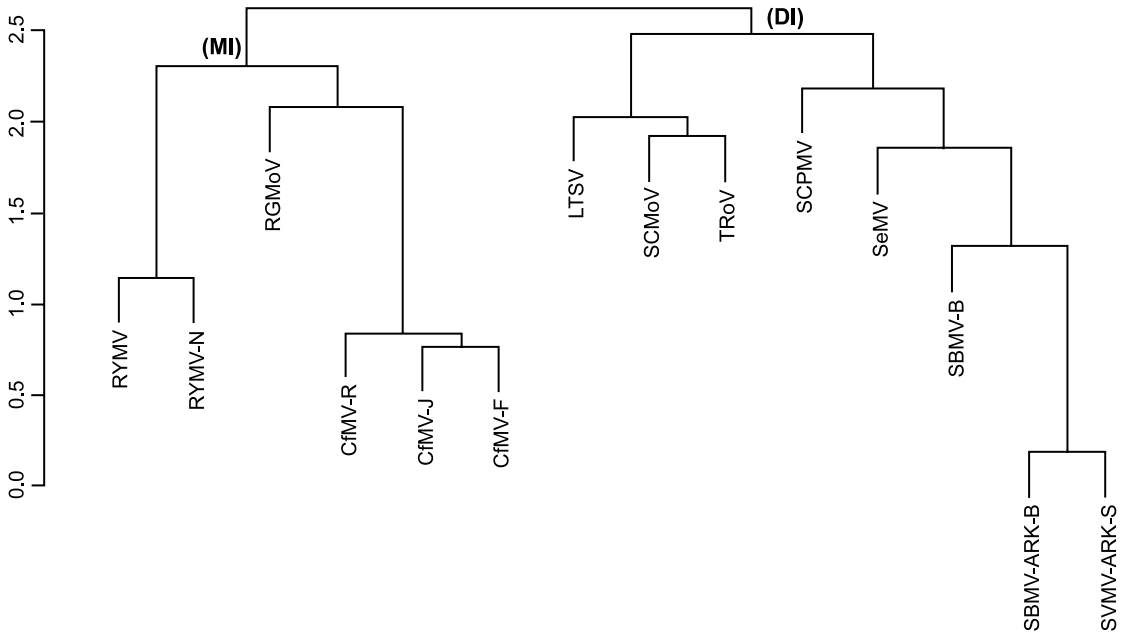

Fig. 2. RSCU relationship tree of sobemoviruses. Dicot- and monocot-infecting viruses were labelled with *DI* and *MI*, respectively. The scale bar showed the *Euclidean distance*

confidence level (42%). Viruses in legumes also broke into two sub-lineages: a SBMV sub-lineage comprising SBMV, SeMV and SCPMV; and a SCMoV sub-lineage comprising SCMoV and LTSV. The *Brassica* virus TRoV was attached to the dicot-infecting lineage. The unrooted ML tree for the genome sequences (Fig. 1B) showed the similar distribution of the viruses, with RGMoV broke away from the other monocot viruses.

Figure 2 shows the RSCU relationship tree of the viruses. RGMoV was clustered with the other two monocot infecting viruses, RYMV and CfMV, suggesting that the RSCU tree appeared more informative than the MP and ML trees (Fig. 1A, B) for sobemoviruses. The dicot-infecting viruses were clustered similarly as in the phylogenetic trees, except that TRoV was in the SCMoV sub-lineage. When the rice and Arabidopsis chromosomes (each treated as an individual species) were added into the RSCU tree (Fig. 3A), the monocot viruses clustered with the rice chromosomes. However, the dicot-infecting SBMV sub-lineage detached from the SCMoV sub-lineage that was grouped with the Arabidopsis chromosomes. SCPMV further mixed into the monocot (M) group. More plant species were then added into the RSCU tree to make 17 species for both dicots and monocots, respectively. With the exception of SCPMV, all dicot-infecting viral species were distant from the dicot (D) cluster, forming a separated dicot-infecting viral cluster that was less distant to monocot species (Fig. 3B).

The tree (Fig. 3B) was robust when more plant species (up to 9 dicots and 2 monocots) were gradually added (data not shown). The stability of the RSCU tree (Fig. 3B) was further tested by removing the 5 monocot species and the 12

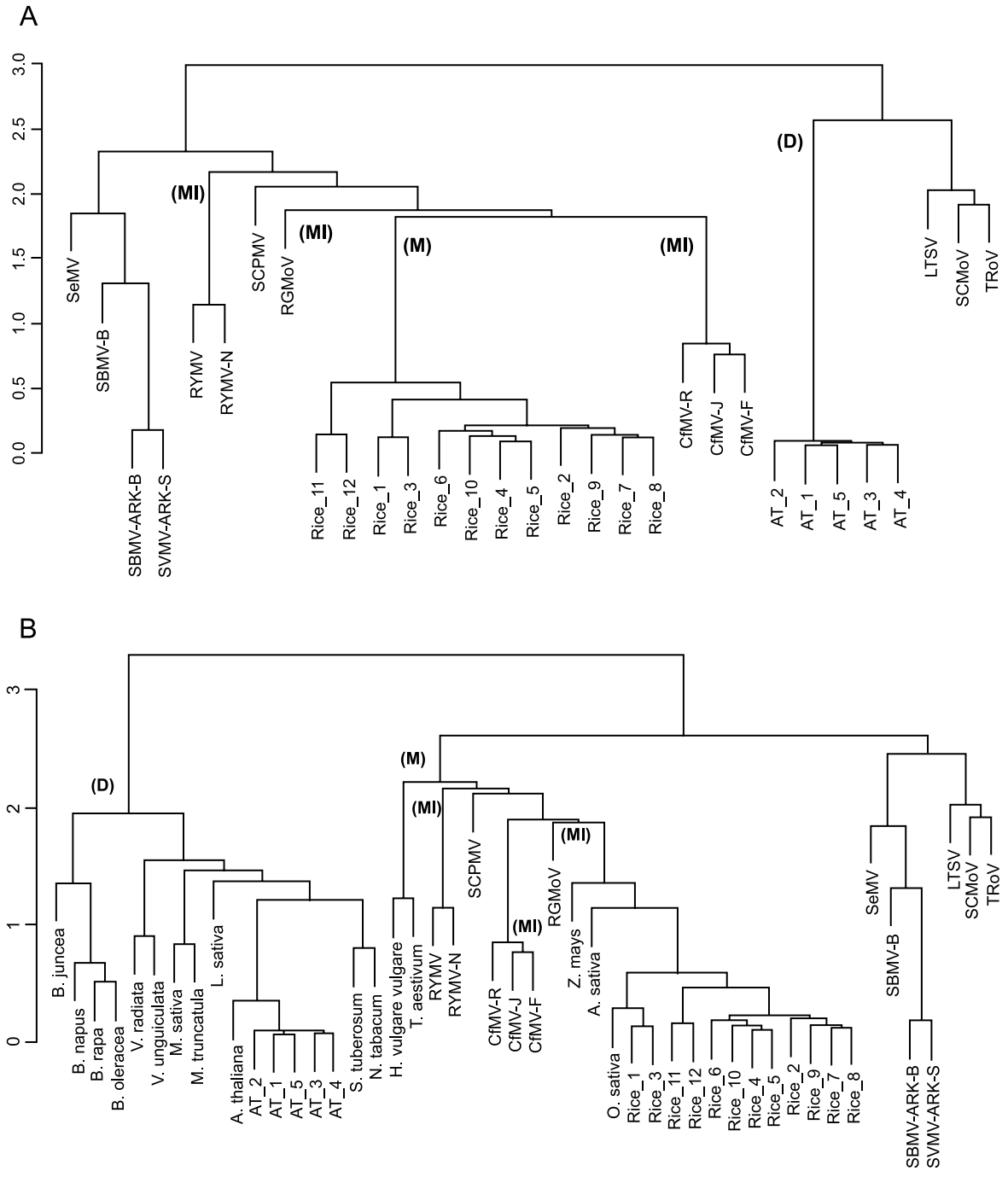

Fig. 3. RSCU relationship trees comparing sobemoviruses with host species. Dicot and monocot branches were labelled with *D* and *M*, respectively. Monocot-infecting viruses were labelled with MI while dicot-infecting viruses were unlabelled. Postfix indicated the chromosome number of host species. **A**: Initial RSCU tree built with sobemoviruses (Table 2) and chromosomes (Table 1) of rice and Arabidopsis (AT). **B**: RSCU tree including more plant species (Table 3)

dicot species (Table 3), respectively. When the 5 monocot species were removed, the tree (Fig. 3B) suffered only minor changes (Fig. 4A). However, when the 12 dicot species were removed, the dicot-infecting cluster (excluding SCPMV)

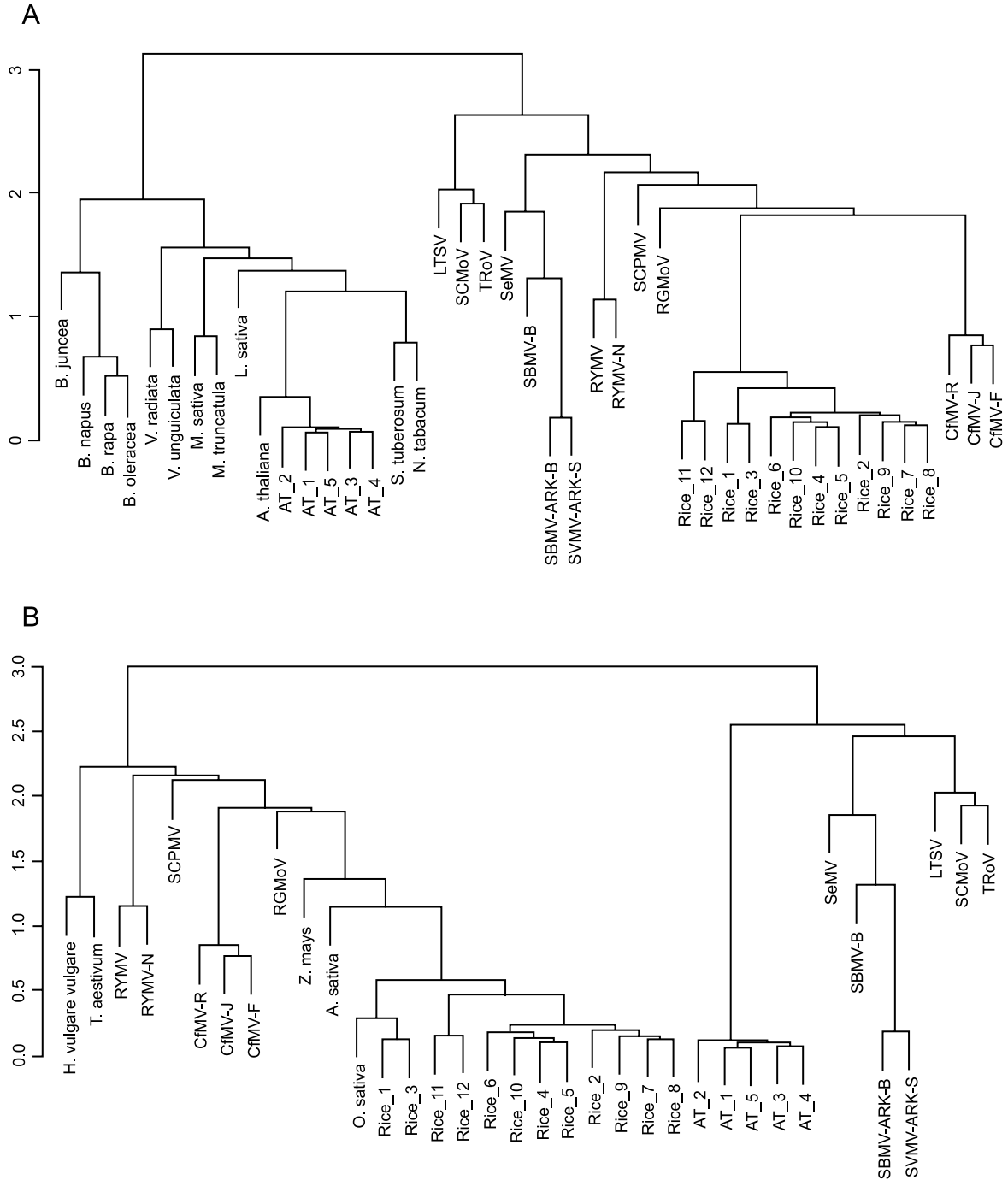

Fig. 4. Stability test of RSCU relationship tree. **A**: Removal of the 5 monocot species (Table 3) from Fig. 3B. **B**: Removal of the 12 dicot species (Table 3) from Fig. 3B

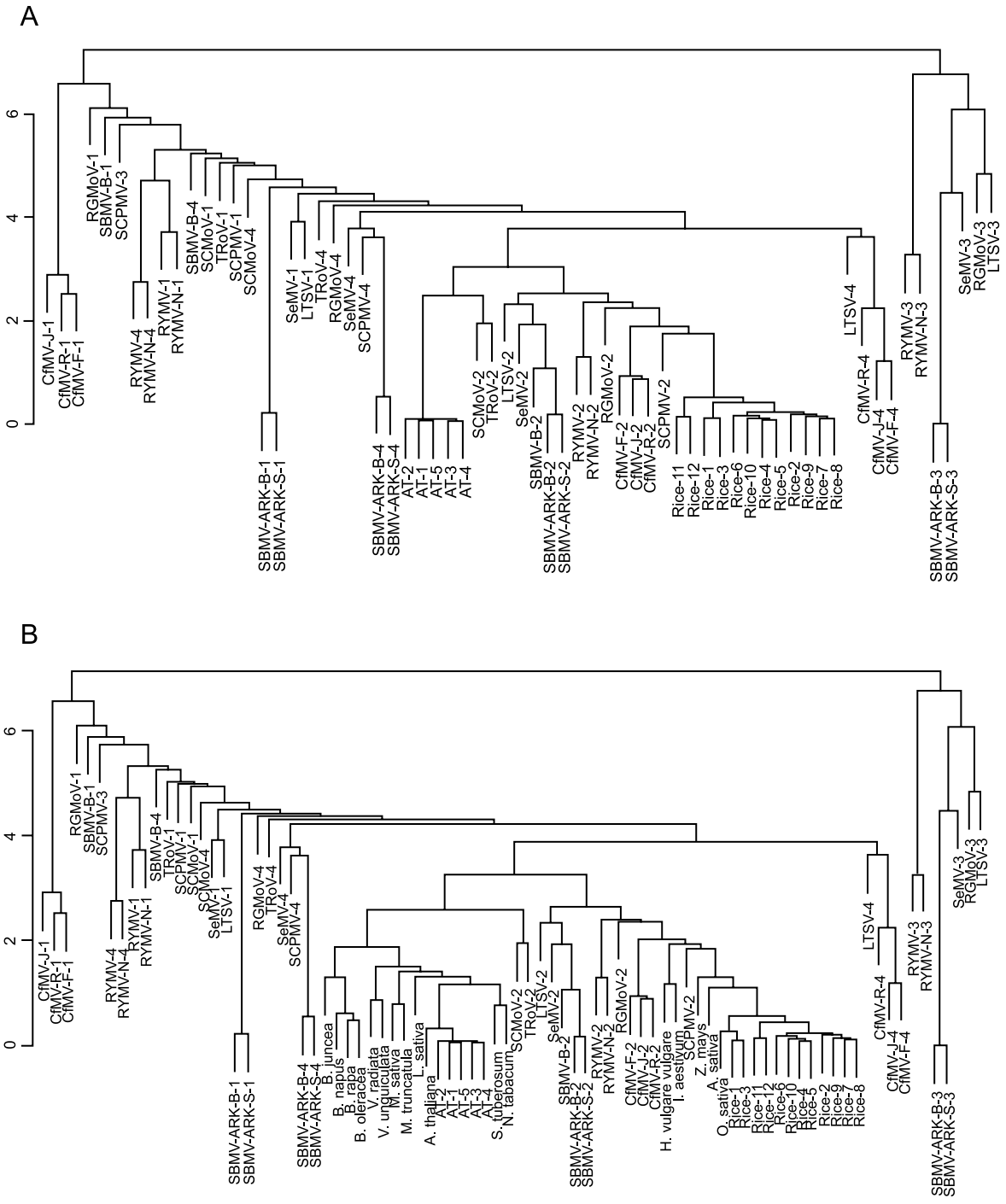

Fig. 5. RSCU relationship trees comparing each ORFs of sobemoviruses with host species. Postfix of rice (Rice) and Arabidopsis (AT) indicated the chromosome numbers. Postfix numbers of viruses indicated the ORF number. **A**: Sobemovirus ORFs and chromosomes of rice and Arabidopsis (AT). **B**: RSCU tree including more (12 dicot and 5 monocot) plant species (Table 3)

shifted to be closer to the Arabidopsis than to the monocot species (Fig. 4B), showing unsatisfactory instability of the RSCU relationship tree (Fig. 3B) when the number of plant species was small.

When a new RSCU relationship tree was then constructed by calculating RSCU for each ORFs of each of the viruses, together with the rice and Arabidopsis data (Fig. 5A), all viral ORF-2s and ORF-3s (except SCPMV) clustered together, respectively. For all viruses, ORF-2s were closer to the ORF-4s than to the ORF-1s. Such relative distance to ORF-2s indicated similarities among ORF-1s and ORF-4s, respectively. Thus, Fig. 5A showed that there was distinguishable codon usage difference among the viral ORFs. Viral ORF-1, -3, and -4 were more distant from the plant chromosomes than that of rice to Arabidopsis, indicating that codon usage bias in these viral ORFs is not attributable to the codon usage bias of plants. Viral ORF-2s were sited within the central lineage containing the plant species, suggesting that the codon usage bias in viral ORF-2s may have genuine affinities with the plant (host) codon usage bias. This relationship tree (Fig. 5A) was more robust than the tree shown in Fig. 3A. When additional plant species were added into the tree (illustrated in Fig. 5B), all viral species retained the positions as in Fig. 5A. Adding more plant (2 monocot and 9 dicot) species did not alter these positions (data not shown). Furthermore, removal of data pertaining to dicot or monocot host species (Table 3) did not cause any alternation of the viral positions (data not shown). Without the noise derived from ORF-1, -3, and -4, all of the monocot-infecting ORF-2s associated with the rice lineage (Fig. 5A), but the rice virus RYMV ORF-2 was the least homologous to rice among the monocot ORF-2s. A dicot virus, SCPMV, was the closest to rice. Four out of six dicot-infecting viruses were more distant to the dicot host cluster than to the monocot cluster (Fig. 5B). Only SCMoV and TRoV were closer to the dicots than to the monocots.

Discussion

Further investigation is needed to reveal why the monocot-infecting RGMoV fell into the dicot-infecting lineages in the genome trees (Fig. 1A, B). Although evidence of recombination among virus species of the sobemovirus genus is lacking [8]. It is possible that the breakaway of RGMoV from the monocot viruses may be due to a recombination event between a dicot and a monocot infecting virus. Viral recombination has been recognised in many +ss RNA viruses, including in *Luteoviridae* [10, 15, 19] in which *Polerovirus* has high homology to the *Sobemovirus* RdRp [6, 16]. However, based on the low confidence level (Fig. 1A, B), a strong case for recombination is not evident. Nevertheless, RGMoV resembled the other monocot-infecting viruses on the bases of their codon usage patterns (Fig. 2), conforming to its monocot-infecting inherence, suggesting that the RSCU tree is perhaps a more informative approach than viral genome trees (Fig. 1A, B) for sobemoviruses.

It is often difficult to compare sequences (nucleotides and amino acids) between viruses and hosts because of low homology. Our approach involving

assembly of RSCU relationships overcomes this problem. The average codon usage pattern (e.g. RSCU table) of any species can be obtained with relative ease as they may have already been available, otherwise RSCU values can be based on encoding sequences. As we demonstrated here, any virus can be compared with a range of susceptible and/or nonsusceptible hosts. Such an analytical capability is advantageous. However, as in all research activities, data (from the codon trees) need to be interpreted sensibly, and with care. For the sobemoviruses we analysed, it is more likely that the sobemoviral ORF-2 originated in a monocot host rather than in a dicot. This is because 4/6 dicot-infecting viral ORF-2s displayed codon usage patterns that resembled those of monocot species, while none of the monocot-infecting ORF-2s showed similarity to the dicot species (Fig. 5A, B). Such disassociation of dicot-infecting viruses to their hosts (Fig. 5A, B) is consistent with the notion that plant viruses do not generally adapt to the host codon usage bias [1]. Though our analysis did not provide definitive proof that sobemoviral ORF-2 originated in monocot hosts rather than in dicot hosts, we anticipate that further analysis of codon usage bias may provide valuable insights into plant virus evolution. Based on the separation between the monocot-infecting cluster and dicot-infecting cluster (Fig. 2), it is tempting to speculate that there may be selection pressure on codon usage bias of sobemoviruses. Furthermore, the RSCU affinities detected between two dicot-infecting viruses (TRoV and SCMoV) and Arabidopsis (Fig. 5A, B), is consistent with the possibility of some degree of positively selected codon adaptation. Similarly, possible negative selection may explain why RYMV showed the least affinity to its current natural host – rice among the three monocot viruses (Fig. 5A, B). Although such speculations might be supported by studies on codon usages in both plant and human viruses that minor selection pressures may act under certain circumstances for either positive or negative selection [1, 5], the rules of selection pressures on viral codon usages, if they exist, are largely unknown. Also, as others have shown [1, 5], mutational bias would play a major role in the emergence of the 'errors' of the RSCU tree. Although in theory such errors should be neutral, however, we could not rule out, at this stage, that the similarity between viral ORF-2 to monocots might be a coincidence. Estimation of confidence for the branches may be needed in future studies.

In sobemoviral genomes, ORF-3 is nested within ORF-2 and is therefore likely to have a common origin with ORF-2, but putatively translated by -1 reading frame shift. This should explain why ORF-3 appeared as an outlier in the RSCU tree (Fig. 5A, B). It is important to note that the putative ORF-3 only exists in 6 out of the 9 sobemoviruses. Sobemovirus ORF-1 encodes a gene silencing suppressor (P1), while ORF-4 is translated from a sub-genomic RNA for the viral CP. All viral ORF-4s were closer to the hosts compared to the distance between their ORF-1s and plants (Fig. 5A, B). This difference in RSCU patterns in sobemovirus genes supports the modular evolution hypothesis that was originally proposed for bacteria phages [2] then considered also applicable to plant viruses (e.g., [4]). These two viral genes may have different origins, and these origins should also differ from that of ORF-2.

Accepting the modular evolution hypothesis, it should be logical that the RSCU tree will become more robust when RSCU values are calculated for each evolutionary module (e.g. ORF) (as shown in Fig. 5A, B) rather than the average of all ORFs (Figs. 3–4). Equally, our approach using the average RSCU values to represent each host species may reduce the accuracy of the trees. The consistent self-clustering of rice chromosomes and Arabidopsis chromosomes indicates that a RSCU tree for host species will be very robust when genome sequences are available for more species. Differences on GC content, codon usage, and amino acid usage between dicots and monocots have been recognised but how such heterogeneities arose remains unclear [17]. It is interesting that all of the four *Brassica* species clustered together (Figs. 3B, 5B) but were distant to Arabidopsis (that also belongs to the family of Brassicaceae). Therefore the RSCU tree approach may also be useful to reveal features of plant genomes, as more become known.

Acknowledgement

HW thanks Dr. I. J. Cooper for critical reading of the manuscript, and NERC (UK) for funding. We appreciate the comments offered by referees.

References

1. Adams MJ, Antoniw JF (2003) Codon usage bias amongst plant viruses. Arch Virol 149: 113–135
2. Botstein D (1980) A theory of modular evolution for bacteriophages. Ann NY Acad Sci 354: 484–491
3. Dwyer GI, Njeru R, Williamson S, Fosu Nyarko J, Hopkins R, Jones RA, Waterhouse PM, Jones MG (2003) The complete nucleotide sequence of Subterranean clover mottle virus. Arch Virol 148: 2237–2247
4. Hull R (2002) Matthews' plant virology, 4th edn. Academic Press, London
5. Jenkins GM, Holmes EC (2003) The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 92: 1–7
6. Koonin EV (1991) The Phylogeny of RNA-Dependent RNA Polymerases of positive-strand RNA viruses. J Gen Virol 72: 2197–2206
7. Lee L, Anderson EJ (1998) Nucleotide sequence of a resistance breaking mutant of southern bea mosaic virus. Arch Virol 143: 2189–2201
8. Lokesh GL, Gopinath K, Satheshkumar PS, Savithri HS (2001) Complete nucleotide sequence of Sesbania mosaic virus: A new virus species of the genus *Sobemovirus*. Arch Virol 146: 209–223
9. Makinen K, Tamm T, Naess V, Truve E, Puurand U, Munthe T, Saarma M (1995) Characterization of cocksfoot mottle sobemovirus genomic RNA and sequence comparison with related viruses. J Gen Virol 76: 2817–2825
10. Moonan F, Molina J, Mirkov TE (2000) Sugarcane yellow leaf virus: An emerging virus that has evolved by recombination between luteoviral and poleroviral ancestors. Virology 269: 156–171
11. Ngon AYM, Ritzenthaler C, Brugidou C, Fauquet C, Beachy RN (1994) Nucleotide sequence and genome characterization of rice yellow mottle virus RNA. J Gen Virol 75: 249–257

12. Othman Y, Hull R (1995) Nucleotide sequence of the bean strain of Southern bean mosaic virus. Virology 206: 287–297
13. Ryabov EV, Krutov AA, Novikov VK, Zheleznikova OV, Morozov SY, Zavriev SK (1996) Nucleotide sequence of RNA from the sobemovirus found in infected cocksfoot shows a luteovirus-like arrangement of the putative replicase and protease genes. Phytopathology 86: 391–397
14. Sharp PM, Li WH (1986) The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281–1295
15. Smith GR, Borg Z, Lockhart BEL, Braithwaite KS, Gibbs MJ (2000) Sugarcane yellow leaf virus: A novel member of the *Luteoviridae* that probably arose by inter-species recombination. J Gen Virol 81: 1865–1869
16. Tamm T, Truve E (2000) Sobemoviruses. J Virol 74: 6231–6241
17. Wong GKS, Wang J, Tao L, Tan J, Zhang J, Passey DA, Yu J (2002) Compositional gradients in *Gramineae* genes. Genome Res 12: 851–856
18. Wu S, Rinehart CA, Kaesberg P (1987) Sequence and organization of Southern bean mosaic virus genomic RNA. Virology 161: 73–80
19. Zanotto PMDA, Gibbs MJ, Gould EA, Holmes EC (1996) A reevaluation of the higher taxonomy of viruses based on RNA polymerases. J Virol 70: 6083–6096

Author's address: Dr. H. Wang, NERC/CEH-Oxford, Mansfield Road, Oxford OX1 3SR, U.K.; e-mail: huw@ceh.ac.uk