DOI 10.1099/vir.0.009266-0

Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses

Siobain Duffy[1,2] and Edward C. Holmes[1,3]

[1]Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA

[2]Department of Ecology, Evolution and Natural Resources, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, NJ 08901, USA

[3]Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA

Correspondence
Siobain Duffy
duffy@aesop.rutgers.edu

Received 30 November 2008
Accepted 17 February 2009

Whitefly-transmitted geminiviruses are major pathogens of the important crop cassava in Africa. The intensive sampling and sequencing of cassava mosaic disease-causing viruses that occurred in the wake of a severe outbreak in Central Africa (1997–2002) allowed us to estimate the rate of evolution of this virus. East African cassava mosaic virus and related species are obligately bipartite (DNA-A and DNA-B segments), and these two genome segments have different evolutionary histories. Despite these phylogenetic differences, we inferred high rates of nucleotide substitution in both segments: mean rates of 1.60×10^{-3} and 1.33×10^{-4} substitutions site^{-1} year^{-1} for DNA-A and DNA-B, respectively. While similarly high substitution rates were found in datasets free of detectable recombination, only that estimated for the coat protein gene (*AV1*), for which an additional DNA-A sequence isolated in 1995 was available, was statistically robust. These high substitution rates also confirm that those previously estimated for the monopartite tomato yellow leaf curl virus (TYLCV) are representative of multiple begomoviruses. We also validated our rate estimates by comparing them with those depicting the emergence of TYLCV in North America. These results further support the notion that geminiviruses evolve as rapidly as many RNA viruses.

INTRODUCTION

Despite intensive study, our understanding of the evolution of plant viruses has lagged behind that of animal viruses. For instance, while many rigorous estimates of rates of nucleotide substitution exist for mammalian RNA viruses (e.g. Chen & Holmes, 2006; Hanada *et al.*, 2004; Jenkins *et al.*, 2002; Ramsden *et al.*, 2008), only in the past year have well-supported estimates of substitution rates of plant viruses been published: for the positive-sense single-stranded RNA (ssRNA+) viruses rice yellow mottle virus (Fargette *et al.*, 2008), zucchini yellow mosaic virus (Simmons *et al.*, 2008), a broad group of potyviruses (Gibbs *et al.*, 2008), and the single-stranded DNA (ssDNA) tomato yellow leaf curl virus (TYLCV; Duffy & Holmes, 2008). Importantly, these data challenge earlier ideas that plant viruses evolve more slowly than animal viruses (Blok *et al.*, 1987; Fraile *et al.*, 1997), as their substitution rates [$\sim5\times10^{-4}$ substitutions site^{-1} year^{-1} (sub site^{-1} year^{-1}) in the capsid proteins] are very similar to those observed in mammalian RNA viruses (Jenkins *et al.*, 2002).

These new estimates also suggest that ssDNA and RNA viruses evolve at similar rates. Because ssDNA viruses replicate using their host cell's DNA polymerases, it has long been thought that ssDNA viruses mutate, and thereby evolve, more slowly than RNA viruses. However, the rates of mutation estimated in the ssDNA bacteriophages ϕX174 [single site reversions $\geqslant1\times10^{-6}$ mutation site^{-1} replication^{-1} (mut site^{-1} rep^{-1}); Denhardt & Silver, 1966; Fersht, 1979; Raney *et al.*, 2004] and M13 (7×10^{-7} mut base^{-1} rep^{-1}; Drake, 1991) are broadly similar to those estimated for the plant RNA virus tobacco mosaic virus (TMV) at 7.26–10.3×10^{-6} mut base^{-1} rep^{-1} (Malpica *et al.*, 2002). Hence, both RNA and ssDNA viruses might be expected to show similar evolutionary dynamics.

To determine whether the high substitution rates observed in TYLCV are widely applicable to other ssDNA plant viruses, we estimated the evolutionary rate of another whitefly-transmitted begomovirus: the East African cassava mosaic disease (CMD)-causing viruses. CMD has a huge economic impact on the staple crop of sub-Saharan Africa (Fauquet & Fargette, 1990; Legg & Fauquet, 2004), and an emergent recombinant of East African cassava mosaic virus

Supplementary material is available with the online version of this paper.

(EACMV) and African cassava mosaic virus (ACMV) destroyed cassava production in Uganda and nearby nations in the late 1990s (Zhou *et al.*, 1997). This led to intensive sampling and sequencing of field isolates of CMD-causing viruses, creating a database of genome sequences sampled over an 8 year period.

Unlike TYLCV, which is frequently monopartite, EACMV and its related species are obligately bipartite, as is more typical of begomoviruses (Gutierrez, 1999). These two circular segments, DNA-A and DNA-B, can reassort amongst genotypes and have distinct evolutionary histories, reflected in non-identical phylogenies (Maruthi *et al.*, 2004; Ndunguru *et al.*, 2005). On the sense strand, the DNA-A encodes two partially overlapping genes for the coat protein (AV1) and a pre-coat protein (AV2, Gutierrez, 1999). In complement, the DNA-A encodes the replication-associated protein (AC1), a transcriptional activator of the coat protein (AC2), a replication enhancer (AC3) and a protein that assists in inter-cell movement (AC4; Gutierrez, 1999). *AC2* and *AC3* partially overlap and *AC4* is entirely encoded in a different reading within *AC1*. The DNA-B segment contains only two non-overlapping genes for BV1: a nuclear shuttle protein involved in export of DNA from the nucleus, and BC1, another protein involved in inter-cell movement (Gutierrez, 1999).

We estimated rates of nucleotide substitution for the entire DNA-A and DNA-B genome segments of four East African CMD-causing viral species (EACMVs), and for individual genes on both segments: *AV1*, *AC1*, *BV1* and *BC1*. This allowed us to assess whether EACMVs evolve at similar rates to TYLCV, and whether there were gene-specific variations in substitution rate, perhaps reflecting functional differences. Unfortunately, we found no evidence of temporal structure in the DNA-B sequences available, nor in the *AC1* dataset, so we were unable to make such comparisons in these cases. We also compared our more robust estimates of substitution rate for the EACMVs' DNA-A and *AV1* gene with epidemiologically determined rates of nucleotide substitution in recently isolated TYLCV.

METHODS

Complete sequences of DNA-A and DNA-B segments of EACMVs with known dates of isolation were downloaded from GenBank (Supplementary Table S1, available in JGV Online). As the DNA-A segment determines which 'species' a geminivirus is, only DNA-B segments isolated with an EACMVs DNA-A were included in this analysis. Further, we only used sequences from viruses that had not been extensively passaged in the laboratory before sequencing, so that the sequences accurately reflected the sequence of the virus in nature. The years of viral isolation (between 1995 and 2002) are given in Supplementary Table S1. The final alignment length and number of taxa in each alignment are given in Table 1.

The circular genome sequences of the whole DNA-A and DNA-B segments were arranged so that each alignment began at the nick-site in the common origin of replication (TAATATT 3′ / 5′AC, Padidam *et al.*, 1995). Datasets were then aligned using MUSCLE (http://www.drive5.com/muscle/, Edgar, 2004), and manually adjusted using Se-Al (http://tree.bio.ed.ac.uk/software/).

Individual gene datasets. Two protein-coding gene alignments were derived from the DNA-A alignment: *AV1* (coat protein gene) and *AC1* (replication-associated protein, which is encoded in the complement to the ssDNA genome), which make up 66 % of the DNA-A. The DNA-A alignment was truncated to create preliminary alignments for these genes. Similarly, the DNA-B alignment was truncated to preliminary alignments for *BV1* (nuclear shuttle protein gene) and *BC1* (movement protein, which is encoded in the complement to the ssDNA genome), which comprise 61 % of the DNA-B. The *BV1* alignment was trimmed by six 5′ codons, because half of the genes began at the seventh codon of *BV1* of the annotated EACMV DNA-B (GenBank accession no. AF112355). This ensures that all of the sequence in each of the individual gene alignments is protein-coding.

Detection of recombination. Putative recombinant genomes were identified using six recombination detection programs within the RDP3 package (http://darwin.uvigo.es/rdp/rdp.html): RDP, GENECONV, MaxChi, Chimaera, Bootscan and 3Seq (Martin *et al.*, 2005). The

Table 1. Details of sequence alignments and parameter estimates

BSP, Bayesian skyline plot; HPD, highest probability density; TMRCA, time to most recent common ancestor.

	Full DNA-A	*AV1*	*AC1*	Full DNA-B	*BV1*	*BC1*
Recombination detected	Yes	No	No	Yes	No	No
Best-fit population growth model	Exponential	Exponential	BSP	BSP	BSP	BSP
Sequence length (nt)	2807	771	1077	2814	774	921
Number of sequences	72	71	39	46	46	33
Time span of sequences	1995–2002	1995–2002	1997–2002	1997–2002	1997–2002	1997–2002
Chain length, in millions	50	15	13	50	20	20
Mean substitution rate	1.60×10^{-3}	1.37×10^{-3}	1.24×10^{-3}	1.33×10^{-4}	2.77×10^{-4}	3.45×10^{-4}
95 % HPD substitution rate	6.13×10^{-4}–2.64×10^{-3}	4.69×10^{-4}–2.32×10^{-3}	6.88×10^{-5}–2.49×10^{-3}	1.06×10^{-5}–3.39×10^{-4}	9.18×10^{-6}–6.46×10^{-4}	6.72×10^{-6}–9.14×10^{-4}
Mean TMRCA (ybp)	71	91	935	896	341	291
95 % HPD TMRCA (ybp)	34–127	34–175	79–2754	90–2592	37–1147	26–1031
d_N/d_S ratio	–	0.11	0.18	–	0.18	0.14

default detection thresholds were employed in all cases. All putative recombinant sequences identified by at least two of these programs were removed from the gene alignments, but not the genomic segment alignments. This was a conservatively low threshold of detectable recombination, ensuring that our results have a high likelihood of being unaffected by recombination. The final numbers of taxa and the length of each alignment are given in Table 1.

Phylogenetic and coalescent analyses. The best-fitting model of nucleotide substitution for each alignment was determined using MODELTEST (Posada & Crandall, 1998). For each dataset, this was the general time-reversible model with invariant sites ($\mathrm{GTR}+\mathrm{I}+\Gamma_4$).

The rates of nucleotide substitution per site and the times to most recent common ancestor (TMRCA) were estimated with the Bayesian Markov chain Monte Carlo (MCMC) method available in the BEAST package (http://beast.bio.ed.ac.uk/, Drummond & Rambaut, 2007). Both strict and relaxed (uncorrelated lognormal) molecular clocks were utilized for each dataset (Drummond *et al.*, 2006). In addition, five demographic models (constant population size, expansion, exponential, and logistic population growth, and a piece-wise Bayesian skyline plot) were used as coalescent priors, although demographic history is a nuisance parameter in this study. We used Bayes factors ($\log_{10}$BF) on tree likelihoods to choose the best-fitting models (in Tracer 1.4, http://beast.bio.ed.ac.uk/Tracer), and assumed a threshold of logBF>2 for significance (http://beast.bio.ed.ac.uk/Model_comparison). Sufficient MCMC chains were run to ensure convergence, with an initial 10 % of the MCMC chains discarded as burn-in. Statistical uncertainty around the mean estimates was provided by the 95 % highest probability density (HPD) values (Table 1). Finally, the BEAST analysis also enabled us to infer maximum clade credibility (MCC) trees for each dataset, in which posterior probability values (a measure of statistical support) were available for each node.

Assessing temporal structure. To test the strength of temporal signal in these data, which is essential to the estimation of substitution rates, we repeated the BEAST analysis (using identical parameters) on a dataset in which sampling times were randomized on the tips (Ramsden *et al.* 2008). The HPDs of these randomized sequences were then compared with those of the real data. If these samples contain clear temporal structure, then the real and resampled data would have different mean estimated substitution rates, and different distributions. As it was not practical to run a large enough number of these resampled runs to garner rigorous statistical comparisons, we compared our results with those of a single resampled run.

Additional sequence analyses. The mean number of non-synonymous (d_N) to synonymous (d_S) nucleotide substitutions per site (d_N/d_S ratio) in each gene alignment was estimated using the single likelihood ancestor counting (SLAC) method, with the appropriate nucleotide substitution model determined by MODELTEST on an initial neighbour-joining phylogenetic tree (Pond & Frost, 2005).

We also employed a phylogenetic approach to detect any mutational bias in the full DNA-A genome dataset. Detailed methods are as published previously (Duffy & Holmes 2008). Briefly, we aligned the DNA-A dataset with its sister taxon [South African cassava mosaic virus (SACMV), GenBank accession number AF155806] as the outgroup (Berrie *et al.*, 2001; Legg & Fauquet, 2004; Maruthi *et al.*, 2004; Ndunguru *et al.*, 2005), and inferred the maximum-likelihood (ML) tree using PAUP* 4.0 (Swofford, 2003), employing tree bisection-reconnection (TBR) branch-swapping. We then tallied the numbers of each kind of substitution and determined deviations from expected values with χ^2 tests of significance. A similar analysis was not conducted for the full DNA-B genome dataset because the published sister taxa aligned very poorly with the dataset.

RESULTS

Frequent recombination in EACMVs' genomes

Frequent recombination, manifest as topological incongruence of different subgenomic regions, was detected in the DNA-A full segment alignment. This included confirmation of the recombinant origins of EACMV-UG (Zhou *et al.*, 1997), which includes EACMV-TZ10 (Ndunguru *et al.*, 2005), EACMV-KE2 (Bull *et al.*, 2006), East African cassava mosaic Zanzibar virus (EACMZV; Maruthi *et al.*, 2004) and African cassava mosaic Kenya virus (EACMKV; Bull *et al.*, 2006), which were all supported by all six methods in RDP (data not shown). The strong monophyly of these species and subspecies can be seen in the MCC tree in Fig. 1. However, the effects of recombination of the entire coat protein gene are evident in the MCC tree for *AV1*, in which EACMV-like coat protein genes were found in some EACMZV isolates, EACMKV-like coat protein genes were found in some EACMV isolates, and the EACMV-KE2 subspecies is no longer monophyletic (Fig. 2). The MCC tree for the smaller number of non-recombinant *AC1* sequences provides limited support for species monophyly (Supplementary Fig. S1, available in JGV Online). Strong evidence of recombination was also detected in 20 of the 46 DNA-B sequences (MCC tree shown in Supplementary Fig. S2), although similar tree topologies were observed in the *BV1* (Supplementary Fig. S3) and *BC1* (Supplementary Fig. S4) datasets. ML trees showed very similar topologies to the MCC trees, with many nodes in the DNA-A and DNA-B trees receiving both strong bootstrap support and high posterior probabilities (data not shown).

For both genomic segments, removing recombinants would have created datasets too small for an analysis of substitution rates. Therefore, with the caveat that recombination violates the assumptions of coalescent-based analyses (although it does not necessarily alter the results), we continued with an analysis of the recombination-loaded whole segment alignments. However, to check the validity of these rate estimates, we pruned detectable recombinants from alignments for each of the four genes. As many of the recombinant breakpoints occurred in the intergenic regions, the number of sequences in each gene alignment was not unduly reduced (Table 1).

Estimation of nucleotide substitution rates

Relaxed molecular clocks described all of the six datasets better than strict molecular clock models, though not significantly for *AC1* and *BV1* (Table 1). There were similar, insignificant differences among the fits of the various demographic models, though exponential population growth models were the best-fit for the DNA-A and *AV1* alignments; the flexible Bayesian skyline models were the best-fit for *AC1* and the three DNA-B alignments (Table 1).

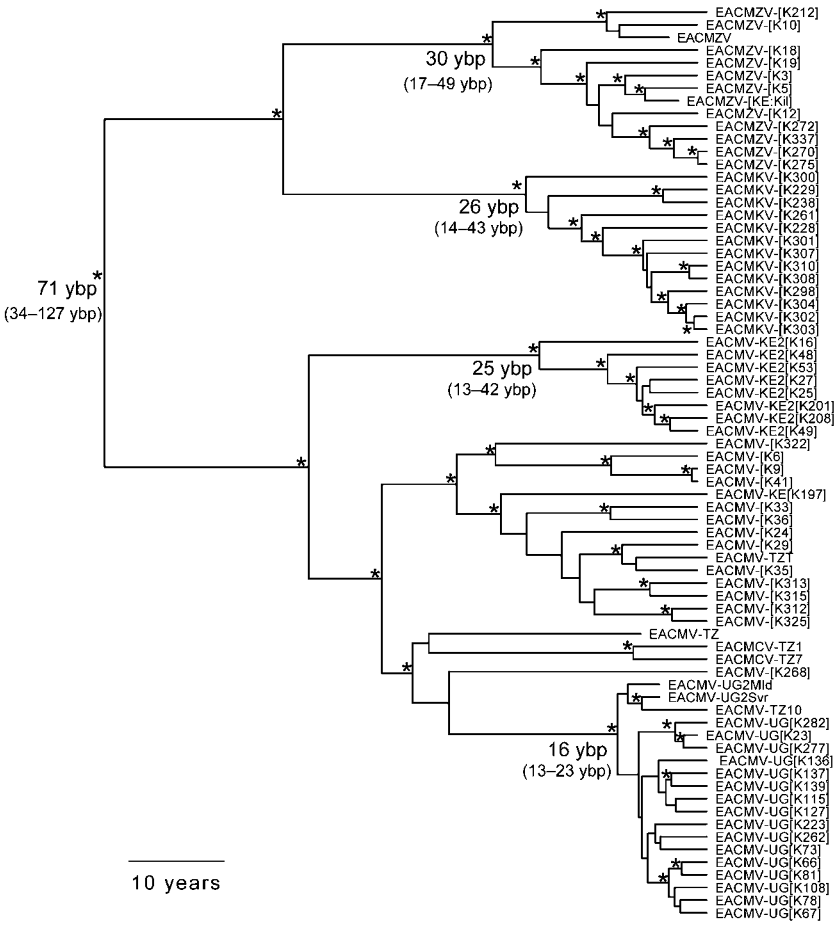

Fig. 1. MCC phylogeny of EACMVs' DNA-As. Branch lengths correspond to the average years of divergence between nodes, with tips given at their years of isolation. The tree is automatically rooted through use of a relaxed molecular clock, and the total depth of the tree is the TMRCA of all sequences (64 years before 2002 – the date of the most recent tip on the tree, 71 ybp). Asterisks denote all nodes with $\geqslant$0.90 posterior probability.

The mean rates of nucleotide substitution estimated for the full genome segments were high, at 1.60×10^{-3} sub site^{-1} year^{-1} (DNA-A, 95 % HPD: 6.13×10^{-4} to 2.64×10^{-3}) and 1.33×10^{-4} sub site^{-1} year^{-1} (DNA-B, 95 % HPD: 1.06×10^{-5} to 3.39×10^{-4}) (Table 1). These rates are within the range documented in RNA viruses (Jenkins *et al.*, 2002), and, for the DNA-A segment, somewhat higher than that estimated for TYLCD-causing viruses (TYLCVs; Duffy & Holmes, 2008)). High rates of nucleotide substitution were also apparent in the four gene datasets that excluded recombinants. The mean substitution rates for *AV1* (1.37×10^{-3} sub site^{-1} year^{-1}) were higher than the other genes (*AC1*=1.24×10^{-3}, *BV1*=2.77×10^{-4}, *BC1*=3.45×10^{-4} sub site^{-1} year^{-1}), though all estimates had overlapping HPDs. These four genes were also shown to be subject to strong purifying selection ($d_N/d_S\leqslant0.18$, Table 1), and there were no positively selected sites in any of the four gene datasets ($P>0.05$). Consequently, these high nucleotide substitution rates are unlikely to result from strong positive selection. The lowest d_N/d_S ratio (0.11) was detected in the gene encoding the coat protein which interacts with both the host and vector, supporting earlier studies (García-Arenal *et al.*, 2001; Seal *et al.*, 2006).

Strength of temporal signal

To determine whether our estimates of substitution rate are robust, and do not simply result from the choice of particular prior values, we randomly reshuffled the dates of isolation for viruses in each of the six datasets and reran BEAST using the best-fitting parameters from the actual analyses. While the HPD intervals of all the resulting reshuffled control datasets overlapped with those of the real data, there were greater differences between the DNA-A and *AV1* analyses and their reshuffled controls. For these two datasets, the reshuffled control 95 % HPD values excluded the mean nucleotide substitution rates estimated from the actual data (Fig. 3). This was not the case for the other four datasets, for which the reshuffled results superimpose on the substitution rates from the real data, indicating that there was insufficient temporal structure to reliably estimate substitution rates. Importantly, these four datasets were sampled over only a 6 year span. In contrast, the DNA-A and *AV1* sequences were sampled over an 8 year span and had more distinct substitution rate estimates from their reshuffled controls, reflecting clear temporal signal (Fig. 3). In addition, this signal was not simply an artefact of the particular reshuffled dates in the control analyses. The average difference between each

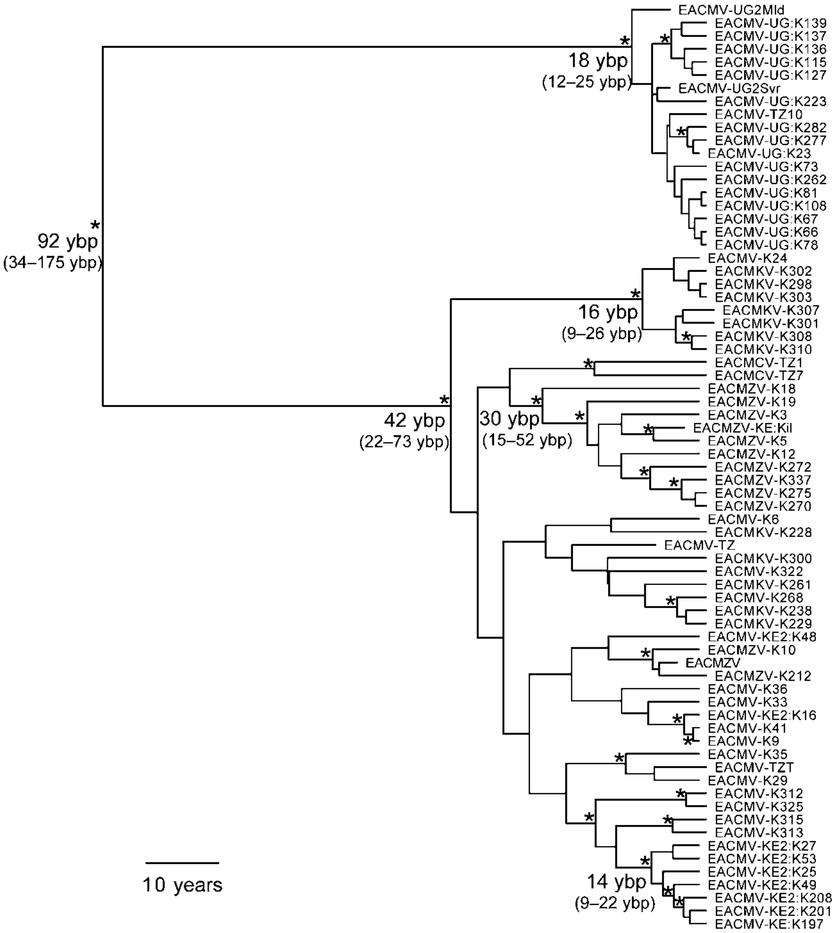

Fig. 2. MCC phylogeny of *AV1* genes from EACMVs. Branch lengths correspond to the average years of divergence between nodes, with tips given at their years of isolation. The tree is automatically rooted through use of a relaxed molecular clock, and the total depth of the tree is the TMRCA of all sequences (84 years before 2002 – the date of the most recent tip on the tree, 91 ybp). Asterisks denote all nodes with ⩾0.90 posterior probability.

reshuffled date and the actual isolation date was similar between datasets with 6 years isolation (0.59, 0.87, 0.83 and 1.15 years for *AC1*, DNA-B, *BV1* and *BC1*, respectively) and those isolated over 8 years (0.86 and 0.93 years for DNA-A and *AV1*, respectively). We concluded that our estimates for the DNA-A and *AV1* were more robust than the others, and that our *AC1* and all of our DNA-B results were potentially unreliable.

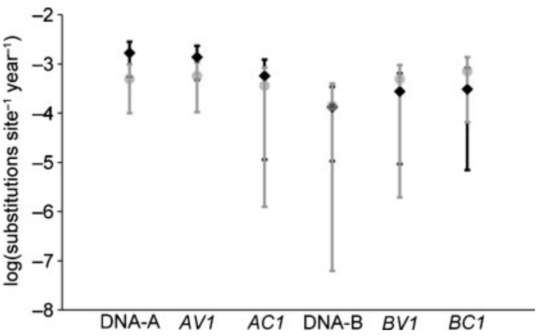

Fig. 3. Average rates of nucleotide substitution of EACMVs estimated from the correct dates of isolation (black diamonds) and the control, reshuffled dates of isolation (grey circles), with 95 % HPD. The DNA-A and *AV1* datasets have dates of isolation ranging from 1995 to 2002; the *AC1*, DNA-B, *BV1* and *BC1* datasets have dates of isolation from 1997 to 2002.

Ages of EACMVs

The DNA-A segments from the four EACMD-causing viral species (EACMVs) have a very short TMRCA [mean= 71 years before present (ybp), 95 % HPD: 34–127 ybp, Table 1]. The detectable-recombinant-free, subgenomic *AV1* dataset produced very similar estimates (TMRCA=91 ybp, 95 % HPD: 23–175 ybp). The remaining major gene on the DNA-A had an older estimated TMRCA of 935 ybp, but the *AC1* analysis was very similar to its reshuffled dates control, and so its estimates are probably not credible. Similarly, as the DNA-B results are not significantly better than a randomly reshuffled date of isolation analysis, we did not consider the TMRCA of these datasets to be reliable.

The DNA-A MCC tree reveals that all completely sequenced isolates of EACMZV share a most recent common ancestor within the last 30 years (95 % HPD: 17–49 ybp, Fig. 1). CMD has been present on Zanzibar since 1933, but has been replaced by the milder EACMZV (Maruthi *et al.* 2004), and our results confirm that the first CMD-causing viruses on this island could not have been EACMZV. EACMKV probably dates back 26 years (95 % HPD: 14–43 ybp). The EACMV-KE2 subspecies is approximately as old (25 years b.p., 95 % HPD: 13–42 ybp), meaning that two major clades of CMD-causing viruses were founded in Kenya at approximately the same time. All the sequences of the severe EACMV-UG variant have a more recent TMRCA of 16 ybp (95 % HPD: 13–23 ybp), or 1993. However, the severe disease symptoms associated with this variant were first observed in north-eastern Uganda in 1988, close to the upper 95 % HPD value of the TMRCA (Zhou *et al.*, 1997). The two dated EACMCV isolates, which were sampled in Tanzania (Ndunguru *et al.*, 2005), also have a recent common ancestor (mean TMRCA of 15 years, 95 % HPD: 10–22 ybp). That these isolates, which differ at ~8 % of sites (Ndunguru *et al.*, 2005), probably diverged within the last 22 years, means there could be a fairly recent TMRCA for the EACMCV isolated on both the Western and Eastern coasts of Africa (~10 % sequence difference). While we cannot provide more conclusive evidence for this without dated sequences of EACMCV from West Africa, our results do call into question the current notion that the Western and Eastern EACMCV isolates 'have been separated for a very long time' (Ndunguru *et al.*, 2005).

Substitution bias in the EACMVs' genome

As EACMVs exhibited RNA virus-like rates of nucleotide substitution in a similar manner to TYLCVs, we tested whether they also show the same substitution biases. After adjusting for the initial nucleotide frequency, six of the 12 substitutions were significantly ($P<0.01$) over- (C→T, G→A, T→G) or under- (A→G, G→T, T→C) represented compared with expectation over the evolution of the DNA-A (Table 2).

Validation of high rates of evolutionary change in begomovirus

The introduction of TYLCV into the New World from the Old World more than 15 years ago (Duffy & Holmes, 2007; Gilbertson *et al.*, 2007) provided a unique opportunity to validate our estimated substitution rates for EACMVs and our prior rates for TYLCVs. We therefore compared recently isolated full-length genomic sequences of TYLCV with the oldest isolate with a full-length genome sequence from the New World (TYLCV-[DO], GenBank accession no. AF024715, isolated in the Dominican Republic in 1994; Nakhla *et al.*, 1994).

The most recently isolated North American TYLCV genomes in GenBank were sampled in Texas in 2006 (TYLCV-US:TX, GenBank accession no. EF110890; Isakeit *et al.*, 2007) and Arizona in 2007 (TYLCV-US:AZ, EF210554; Idris *et al.*, 2007). Both these sequences were too recent to have been included in the dataset from which our TYLCVs substitution rates were originally estimated (Duffy & Holmes, 2008), and therefore offer an independent measure with which to validate these estimates. However, they are more closely related to each other than to other North American isolates (99.3 % nt identity, Idris *et al.*, 2007), and therefore the calculated substitution rates from these two isolates are not independent. Crucially, the intergenic region and coat protein gene of TYLCV-US:TX, and by extension TYLCV-US:AZ, are directly descended from those of TYLCV-[DO] (i.e. this sequence falls at the node of the tree; Duffy & Holmes, 2007), which is necessary for such an analysis to be accurate.

There were 25 sites that differed between TYLCV-[DO] and TYLCV-US:TX and 22 between TYLCV-[DO] and TYLCV-US:AZ. Of these, 15 were shared by TYLCV-US:TX and TYLCV-US:AZ, including a deletion of 29 bases in the large intergenic region. Each recent isolate had three substitutions in the 777 nt of the coat protein-coding gene, and

Table 2. Nucleotide substitutions inferred from the outgroup-rooted ML phylogeny of EACMVs' DNA-A, evaluated with χ^2 tests of significance

Original base	Resulting base			
	A	C	G	T
A	–	53 (59.3)	151 (201.4)	73 (93.4)
χ^2		0.67	12.6	4.44
P		0.42	3.9×10^{-4}	0.04
C	53 (46.7)	–	69 (57.7)	338 (256.6)
χ^2	0.84		2.21	25.8
P	0.36		0.14	3.8×10^{-7}
G	252 (201.6)	62 (73.3)	–	242 (354.1)
χ^2	12.6	1.74		35.5
P	3.9×10^{-4}	0.19		2.57×10^{-9}
T	122 (101.6)	273 (354.4)	497 (384.9)	–
χ^2	4.08	18.7	32.6	
P	0.04	1.6×10^{-5}	1.1×10^{-8}	

treating the 29 base indel as a single substitution, there were five and four substitutions in the 285 nt of the large intergenic region of the recent isolates, respectively. The differences between these genomes, over the 12 or 13 years separating the sampling of the genomes, represent an evolutionary rate of ~3×10^{-4} for the coat protein-coding gene and one of ~1×10^{-3} for the intergenic region (Fig. 4). Strikingly, there was very good agreement between our earlier estimates for the rate of TYLCV evolution and the substitution rate of TYLCV in North America inferred here. However, our estimated rate of EACMV evolution was higher than that inferred for the DNA-A and coat protein gene of North American TYLCV.

DISCUSSION

Our analysis revealed that DNA-A of EACMVs, including the detectable-recombinant-free coat protein gene, *AV1*, had a substitution rate that is similar to both TYLCVs and RNA viruses. However, we could not draw any definitive conclusions about the evolutionary rate of the DNA-B because our dataset contained viral sequences sampled over too short a time span. EACMV and TYLCV are currently the only begomoviruses with sufficient dated sequences to infer substitution rates, although even these estimates were compromised in datasets sampled across only 6 years. Clearly, a far larger sample of viruses is required to increase the accuracy of all EACMVs rate estimates.

This short time span also likely explains why our estimated nucleotide substitution rates for EACMVs were somewhat higher than for TYLCVs, and why our mean TMRCA for the EACMV-UG clade is shorter than implied by epidemiological data. Sequences sampled over shorter timescales tend to produce increased rate estimates, as they are more likely to reflect the background rate of mutation, prior to the action of purifying selection (i.e. they include transient deleterious mutations), than sequences sampled over longer time periods which produce a better estimate of the true substitution rate (Duffy *et al.* 2008; Holmes, 2003). Indeed, the higher rates in EACMVs are unlikely to be explained by the alternative explanation of adaptive evolution, as there was no evidence for positive selection in these data.

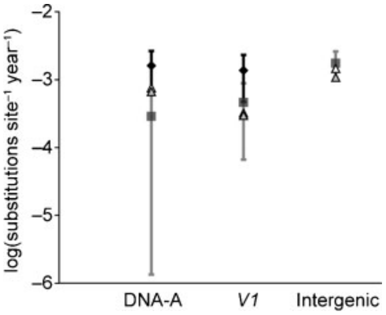

Fig. 4. Estimated rates of nucleotide substitution of TYLCVs and EACMVs. The estimated mean substitution rates for TYLCVs (Duffy & Holmes, 2008), with 95 % HPD values, are shown in dark grey. The mean estimated rates for the DNA-A and *AV1* from EACMVs, with 95 % HPD, are shown in black. The epidemiologically estimated substitution rates between TYLCV-[DO] and TYLCV-US:TX (open triangles) and TYLCV-US:AZ (grey triangles) are also plotted. The full genome of TYLCV is analogous to the DNA-A segment of EAMCV. The *V1* gene encodes the coat protein and is analogous to *AV1* and the intergenic region spans ~300 nt between the start codons of genes for the pre-coat protein (*V2*) and the replication protein (*C1*).

Cassava mosaic disease was first reported in 1894, in Tanzania (Legg & Fauquet, 2004), and is thought to have originated from indigenous African geminiviruses that changed hosts to infect this introduced New World crop (Bull *et al.*, 2006; Fauquet & Fargette, 1990). There are at least five related, geographically grouped clades of CMD-causing viruses: the Indian and Sri Lankan species are more related to each other than to African, East African and South African cassava mosaic viruses (Berrie *et al.*, 2001; Maruthi *et al.*, 2004; Ndunguru *et al.*, 2005; Zhou *et al.*, 1998). Unfortunately, there are very few published dates of isolation for Indian cassava mosaic virus (ICMV), Sri Lankan cassava mosaic virus (SLCMV) ACMV and SACMV sequences, so we could not explore when these clades diverged from one another. However, our mean estimate of 91 years ago (HPD: 34–175 years ago) for the TMRCA of the *AV1* gene of EACMVs is consistent with the emergence of CMD in the nineteenth century, and implies that EACMD-causing viruses may be derived from another species of CMD-causing virus that arose earlier. Similarly, our analyses provide evidence against the idea that Eastern and Western African EACMV strains had diverged prior to the arrival of cassava to Africa in the sixteenth century (Ndunguru *et al.*, 2005).

The measured rate of nucleotide substitution of TYLCV as it moved throughout North America validated our high, RNA virus-like, estimates of substitution rates in two begomoviruses (Fig. 4). While our estimated rates of DNA-A evolution were potentially influenced by recombination, the validation of the rate of evolutionary change in the coat protein gene provides strong evidence that begomovirus genes are indeed evolving rapidly. Similarly, the measured substitution rate of TYLCV validated that the large intergenic region is evolving an order of magnitude more quickly than the coat protein-coding gene, as expected given weaker purifying selection. No analyses of intergenic regions for EACMVs were reported because the two datasets only contained sequences from a 6 year span, and were not different from their reshuffled date controls (data not shown), and therefore were not considered reliable.

The high mutation frequency of geminiviruses (Arguello-Astorga *et al.*, 2007; Ge *et al.*, 2007; Isnard *et al.*, 1998; Shepherd *et al.*, 2005), the high estimated substitution rates of geminiviruses (Duffy & Holmes, 2008; van der Walt *et* *al.*, 2008), and the high mutation frequencies and nucleotide substitution rates of other families of ssDNA viruses (Shackelton & Holmes, 2006; Shackelton *et al.*, 2005; Umemura *et al.*, 2002) necessitate that ssDNA viruses have high mutation rates. The mutation frequencies and rates of the ssDNA phages ϕX174 (Denhardt & Silver, 1966; Raney *et al.*, 2004) and M13 (Drake, 1993) approach that of tobacco mosaic virus ($7.3–10.3 \times 10^{-6}$ mut base^{-1} replication^{-1}; Malpica *et al.*, 2002). If all ssDNA viruses mutate at similar rates, geminiviruses should have mutation rates nearly as high as plant RNA viruses.

It is yet not known how geminiviruses, or any other ssDNA virus, achieve the high mutation rates necessary to explain their genetic variability and their high nucleotide substitution rates while replicating with their hosts' DNA polymerases (Duffy *et al.*, 2008). Possible explanations include the use of little-used, error-prone plant DNA polymerases, or that the geminivirus mutation rate reflects the polymerase mutation rate without subsequent correction by DNA repair enzymes (Duffy & Holmes, 2008), although there is, as yet, no evidence to support either notion. However, the biased substitution patterns found in EACMVs indicate a potential source of mutation besides host polymerase error.

Two of the three significantly overrepresented substitutions from this study (C⟶T and G⟶A) were also over-represented over the evolution of the TYLCVs (Duffy & Holmes, 2008), and are the exact transitions associated with deamination of nucleotide bases (Caulfield *et al.*, 1998). While specific cellular enzymes are known to deaminate bases of viral RNA and DNA (Walsh & Xu, 2006), oxidative deamination can occur spontaneously, and is up to 100 times more likely to occur on ssDNA compared with dsDNA (Frederico *et al.*, 1990). As geminivirus genomes spend time in single-stranded states, they would be susceptible to these spontaneous chemical reactions, and probably do not have access to host DNA-repair enzymes that would correct the damaged, deaminated bases. Evidence for another kind of mutagenic oxidative damage, causing overrepresentation of G⟶T substitutions, has been found on the virion strand of the mastrevirus maize streak virus (van der Walt *et al.*, 2008) and TYLCVs (Duffy & Holmes, 2008). This was not the case in this study, where the reverse T⟶G transversion was over-represented. While we have no biochemical explanation for this over-represented substitution, this pattern has been supported by a recent experimental paper that detected only transition mutations and a single T⟶G transversion in EACMZV (Bull *et al.*, 2007). Although which substitutions are overrepresented may vary among geminiviruses, spontaneous oxidative mutations could be adding to the polymerase-induced mutation rate in all geminiviruses. This mechanism can help to explain the difference between the low estimated per base mutation rate of plant DNA polymerases (Drake *et al.*, 1998) and the necessarily high per base mutation rate of geminiviruses.

ACKNOWLEDGEMENTS

We would like to thank David Robinson, Scottish Crop Research Institute, Dundee, UK, for providing details of the isolation of EACMV-TZ. This work was supported by a National Science Foundation grant DBI-0630707 to S. D.

REFERENCES

Arguello-Astorga, G., Ascencio-Ibáñez, J. T., Dallas, M. B., Orozco, B. M. & Hanley-Bowdoin, L. (2007). High-frequency reversion of geminivirus replication protein mutants during infection. *J Virol* **81**, 11005–11015.

Berrie, L. C., Rybicki, E. P. & Rey, M. E. C. (2001). Complete nucleotide sequence and host range of *South African cassava mosaic virus*: further evidence for recombination amongst begomoviruses. *J Gen Virol* **82**, 53–58.

Blok, J., Mackenzie, A., Guy, P. & Gibbs, A. (1987). Nucleotide sequence comparisons of *Turnip yellow mosaic virus* isolates from Australia and Europe. *Arch Virol* **97**, 283–295.

Bull, S. E., Briddon, R. W., Sserubombwe, W. S., Ngugi, K., Markham, P. G. & Stanley, J. (2006). Genetic diversity and phylogeography of cassava mosaic viruses in Kenya. *J Gen Virol* **87**, 3053–3065.

Bull, S. E., Briddon, R. W., Sserubombwe, W. S., Ngugi, K., Markham, P. G. & Stanley, J. (2007). Infectivity, pseudorecombination and mutagenesis of Kenyan cassava mosaic begomoviruses. *J Gen Virol* **88**, 1624–1633.

Caulfield, J. L., Wishnok, J. S. & Tannenbaum, S. R. (1998). Nitric oxide-induced deamination of cytosine and guanine in deoxynucleosides and oligonucleotides. *J Biol Chem* **273**, 12689–12695.

Chen, R. & Holmes, E. C. (2006). Avian influenza virus exhibits rapid evolutionary dynamics. *Mol Biol Evol* **23**, 2336–2341.

Denhardt, D. T. & Silver, R. B. (1966). An analysis of the clone size distribution of ΦX174 mutants and recombinants. *Virology* **30**, 10–19.

Drake, J. W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci U S A* **88**, 7160–7164.

Drake, J. W. (1993). Rates of spontaneous mutation among RNA viruses. *Proc Natl Acad Sci U S A* **90**, 4171–4175.

Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics* **148**, 1667–1686.

Drummond, A. J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**, e88.

Duffy, S. & Holmes, E. C. (2007). Multiple introductions of the Old World begomovirus *Tomato yellow leaf curl virus* into the New World. *Appl Environ Microbiol* **73**, 7114–7117.

Duffy, S. & Holmes, E. C. (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus *Tomato yellow leaf curl virus* (TYLCV). *J Virol* **82**, 957–965.

Duffy, S., Shackelton, L. A. & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**, 267–276.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797.

Fargette, D., Pinel, A., Rakotomalala, M., Sangu, E., Traoré, O., Sérémé, D., Sorho, F., Issaka, S., Hébrard, E. & other authors (2008). *Rice yellow mottle virus*, an RNA plant virus, evolves as rapidly as most RNA animal viruses. *J Virol* **82**, 3584–3589.

Fauquet, C. & Fargette, D. (1990). African cassava mosaic virus: etiology, epidemiology, and control. *Plant Dis* **74**, 404–411.

Fersht, A. R. (1979). Fidelity of replication of phage ΦX174 DNA by DNA polymerase III holoenzyme: spontaneous mutation by misincorporation. *Proc Natl Acad Sci U S A* **76**, 4946–4950.

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. J Virol* **71**, 8316–8320.

Frederico, L. A., Kunkel, T. A. & Shaw, B. R. (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* **29**, 2532–2537.

García-Arenal, F., Fraile, A. & Malpica, J. M. (2001). Variability and genetic structure of plant virus populations. *Annu Rev Phytopathol* **39**, 157–186.

Ge, L., Zhang, J., Zhou, X. & Li, H. (2007). Genetic structure and population variability of tomato yellow leaf curl China virus. *J Virol* **81**, 5902–5907.

Gibbs, A. J., Ohshima, K., Phillips, M. J. & Gibbs, M. J. (2008). The prehistory of potyviruses: their initial radiation was during the dawn of agriculture. *PLoS One* **3**, e2523.

Gilbertson, R. L., Rojas, M. R., Kon, T. & Jaquez, J. (2007). Introduction of Tomato yellow leaf curl virus into the Dominican Republic: the development of a successful integrated pest management strategy. In *Tomato Yellow Leaf Curl Virus Disease*, pp. 279–303. Edited by H. Czosnek. Dordrecht: Springer.

Gutierrez, C. (1999). Geminivirus DNA replication. *Cell Mol Life Sci* **56**, 313–329.

Hanada, K., Suzuki, Y. & Gojobori, T. (2004). A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Mol Biol Evol* **21**, 1074–1080.

Holmes, E. C. (2003). Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. *J Virol* **77**, 11296–11298.

Idris, A. M., Guerrero, J. C. & Brown, J. K. (2007). Two distinct isolates of *Tomato yellow leaf curl virus* threaten tomato production in Arizona and Sonora, Mexico. *Plant Dis* **91**, 910.

Isakeit, T., Idris, A. M., Sunter, G., Black, M. C. & Brown, J. K. (2007). *Tomato yellow leaf curl virus* in tomato in Texas, originating from transplant facilities. *Plant Dis* **91**, 466.

Isnard, M., Granier, M., Frutos, R., Reynaud, B. & Peterschmitt, M. (1998). Quasispecies nature of three maize streak virus isolates obtained through different modes of selection from a population used to assess response to infection of maize cultivars. *J Gen Virol* **79**, 3091–3099.

Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. (2002). Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol* **54**, 156–165.

Legg, J. P. & Fauquet, C. M. (2004). Cassava mosaic geminiviruses in Africa. *Plant Mol Biol* **56**, 585–599.

Malpica, J. M., Fraile, A., Moreno, I., Obies, C. I., Drake, J. W. & Garcia-Arenal, F. (2002). The rate and character of spontaneous mutation in an RNA virus. *Genetics* **162**, 1505–1511.

Martin, D. P., Williamson, C. & Posada, D. (2005). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21**, 260–262.

Maruthi, M. N., Seal, S., Colvin, J., Briddon, R. W. & Bull, S. E. (2004). *East African cassava mosaic Zanzibar virus* – a recombinant begomovirus species with a mild phenotype. *Arch Virol* **149**, 2365–2377.

Nakhla, M. K., Maxwell, D. P., Martinez, R. T., Carvalho, M. G. & Gilbertson, R. L. (1994). Widespread occurrence of the eastern Mediterranean strain of tomato yellow leaf curl geminivirus in tomatoes in the Dominican Republic. *Plant Dis* **78**, 926.

Ndunguru, J., Legg, J. P., Aveling, T. A. S., Thompson, G. & Fauquet, C. M. (2005). Molecular biodiversity of cassava begomoviruses in Tanzania: evolution of cassava geminiviruses in Africa and evidence for East Africa being a center of diversity of cassava geminiviruses. *Virol J* **2**, 21.

Padidam, M., Beachy, R. N. & Fauquet, C. M. (1995). Classification and identification of geminiviruses using sequence comparisons. *J Gen Virol* **76**, 249–263.

Pond, S. L. K. & Frost, S. D. W. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21**, 2531–2533.

Posada, D. & Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.

Ramsden, C., Melo, F. L., Figueiredo, L. M., Holmes, E. C. & Zanotto, P. M. (2008). High rates of molecular evolution in hantaviruses. *Mol Biol Evol* **25**, 1488–1492.

Raney, J. L., Delongchamp, R. R. & Valentine, C. R. (2004). Spontaneous mutant frequency and mutation spectrum for gene *A* of ΦX174 grown in *E. coli. Environ Mol Mutagen* **44**, 119–127.

Seal, S. E., van den Bosch, F. & Jeger, M. J. (2006). Factors influencing begomovirus evolution and their increasing global significance: implications for sustainable control. *Crit Rev Plant Sci* **25**, 23–46.

Shackelton, L. A. & Holmes, E. C. (2006). Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. *J Virol* **80**, 3666–3669.

Shackelton, L. A., Parrish, C. R., Truyen, U. & Holmes, E. C. (2005). High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proc Natl Acad Sci U S A* **102**, 379–384.

Shepherd, D. N., Martin, D. P., McGivern, D. R., Boulton, M. I., Thomson, J. A. & Rybicki, E. P. (2005). A three-nucleotide mutation altering the *Maize streak virus* Rep pRBR–interaction motif reduces symptom severity in maize and partially reverts at high frequency without restoring pRBR–Rep binding. *J Gen Virol* **86**, 803–813.

Simmons, H. E., Holmes, E. C. & Stephenson, A. G. (2008). Rapid evolutionary dynamics of zucchini yellow mosaic virus. *J Gen Virol* **89**, 1081–1085.

Swofford, D. L. (2003). PAUP* Phylogenetic analysis using parsimony (and other methods), version 4.0b8. Sunderland, MA: Sinauer Associates.

Umemura, T., Tanaka, Y., Kiyosawa, K., Alter, H. J. & Shih, J. W.-K. (2002). Observation of positive selection within hypervariable regions of a newly identified DNA virus (SEN virus). *FEBS Lett* **510**, 171–174.

van der Walt, E., Martin, D. P., Varsani, A., Polston, J. E. & Rybicki, E. P. (2008). Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. *Virol J* **5**, 104.

Walsh, C. P. & Xu, G. L. (2006). Cytosine methylation and DNA repair. *Curr Top Microbiol Immunol* **301**, 283–315.

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

Zhou, X., Robinson, D. J. & Harrison, B. D. (1998). Types of variation in DNA-A among isolates of East African cassava mosaic virus from Kenya, Malawi and Tanzania. *J Gen Virol* **79**, 2835–2840.