Rapid evolutionary dynamics of zucchini yellow mosaic virus

Heather E. Simmons,1 Edward C. Holmes1,2 and Andrew G. Stephenson1

1Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA
2Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA

Zucchini yellow mosaic virus (ZYMV) is an economically important virus of cucurbit crops. However, little is known about the rate at which this virus has evolved within members of the family Cucurbitaceae, or the timescale of its epidemiological history. Herein, we present the first analysis of the evolutionary dynamics of ZYMV. Using a Bayesian coalescent approach we show that the coat protein of ZYMV has evolved at a mean rate of 5.0×10^{-6} nucleotide substitutions per site, per year. Notably, this rate is equivalent to those observed in animal RNA viruses. Using the same approach we show that the lineages of ZYMV sampled here have an ancestry that dates back no more than 800 years, suggesting that human activities have played a central role in the dispersal of ZYMV. Finally, an analysis of phylogeographical structure provides strong evidence for the in situ evolution of ZYMV within individual countries.

Zucchini yellow mosaic virus (ZYMV), first isolated in 1973 and described in 1981 (Lisa et al., 1981), is the cause of one of the most economically important diseases of the family Cucurbitaceae, naturally infecting plants in more than 50 countries (Desbiez & Lecoq, 1997). Symptoms include yellowing, stunting, leaf deformations, and misshaped and discoloured fruits, which often renders the fruits unmarketable, drastically reducing agricultural yields (Blua & Perring, 1989; Desbiez & Lecoq, 1997; Gal-On, 2007). Although ZYMV is widespread, few viral reservoirs have been identified, particularly in temperate regions (Desbiez & Lecoq, 1997).

ZYMV is a single-stranded, positive-sense RNA virus of the family Potyviridae. The primary mode of transmission is via aphids in a non-persistent manner. Although 10 aphid species have been reported as vectors (Katis et al., 2006; Lisa et al., 1981), a wider range of potential aphid vectors has been identified under experimental conditions (Blackman & Eastop, 2000; Katis et al., 2006). While aphid transmission is undoubtedly the main route of spread for ZYMV, infrequent seed transmission has also been proposed (Robinson et al., 1993; Schrijnwerkers et al., 1991), the epidemiological importance of which is uncertain (Johansen et al., 1994).

ZYMV has a genome of 9593 nt arranged as a single open reading frame encoding a polyprotein precursor that is processed into 10 putative proteins (Gal-On, 2007). Of these, the coat protein (CP) is involved in the encapsidation of viral RNA, vector transmission (Shukla et al., 1991; Urcuqui-Inchima et al., 2001), the regulation of viral RNA amplification and cell-to-cell and systemic movement (Urcuqui-Inchima et al., 2001). Transmission occurs as a result of the interaction between the aphid stylet, CP and the HC-Pro protein (Pirone & Blanc, 1996), such that some mutations in CP and HC-Pro disrupt viral transmission (Gal-On, 2007; Pirone & Blanc, 1996; Shukla et al., 1991; Urcuqui-Inchima et al., 2001). The CP is also extensively used as a tool to infer the phylogenetic relationships among viral isolates (Rybicki & Shukla, 1992; Shukla et al., 1991).

A variety of studies have explored the extent and structure of genetic diversity in ZYMV, particularly within a biogeographical context. Analysis of a 250 nt fragment of 160 viral isolates sampled from 23 geographical areas revealed two major groups of ZYMV, denoted A and B, with the former divided into three clusters (Desbiez et al., 2002). A subsequent analysis of the CP revealed three main groups of isolates with differing geographical distributions (Zhao et al., 2003). Group I included the majority of European isolates, as well as some from China and Japan, and a single Californian isolate. Group II was exclusively composed of viruses from Asia, while group III included several Chinese isolates. Notably, while group I and II isolates resulted in mosaic symptoms on leaves and fruit distortion, group III viruses did not cause symptoms on the fruit, but induced severe mosaic symptoms on the leaves (Zhao et al., 2003). More phylogenetically distant ZYMV isolates were observed in Singapore and Réunion (and other islands in the Indian Ocean representing group B of Desbiez et al., 2002), which probably reflects their biogeographical separation (Gal-On, 2007; Zhao et al., 2003).
More localized phylogeographical studies have revealed that viruses can diffuse within specific localities, such as Central Europe (Glasa & Pittnerova, 2006; Glasa et al., 2007; Tobias & Palkovics, 2003), perhaps mediated by the local spread of aphids. However, isolates sampled from adjoining locations are not always related (Plosser & Baumann, 2002), suggesting that biogeographical structure may, to some extent, be determined by the international trading of infected seeds (Desbiez et al., 2002; Tobias & Palkovics, 2003).

There has also been considerable interest in using sequence data from plant RNA viruses to infer evolutionary dynamics. Although a combination of intrinsically high rates of mutation, rapid replication and large population sizes are thought to provide RNA viruses with abundant genetic variation, some plant RNA viruses appear more genetically stable than their animal counterparts (García-Arenal et al., 2001, 2003). This could be due to a combination of intrinsically lower rates of mutation (Malpica et al., 2002) and a reduced fixation rate of advantageous non-synonymous mutations because of weaker immune selection (García-Arenal et al., 2001). Similarly, genetic bottlenecks play a major role in structuring genetic diversity during both systemic infection (French & Stenger, 2003; Li & Roossinck, 2004; Sacristán et al., 2003) and horizontal transmission by aphids (Ali et al., 2006).

Despite the agricultural importance of ZYMV, there has been little work documenting either the rate of molecular evolution of this virus or the age of the sampled genetic diversity, reflected in the time to the most recent common ancestor (TMRCA). However, this information is central to understanding the evolutionary dynamics of plant RNA viruses in general, and particularly whether they exhibit reduced rates of evolutionary change, which in turn may have major implications on their ability to emerge in new host species.

*Cucurbita pepo* ssp. *texana* is an annual monococious vine that is native to northern Mexico, Texas, and the lower Mississippi River drainage area. It is thought to be either the wild progenitor of the cultivated squashes (*C. pepo* ssp. *pepo*) or an early escape from cultivation (Decker & Wilson, 1987; Decker-Walters, 1990; Decker-Walters et al., 2002; Lira et al., 1995). ZYMV infection of plants collected during the 2006 growing season was determined immunologically (DAS-ELISA test kit; Agdia). Leaf tissue from infected plants was then homogenized in liquid nitrogen and RNA extracted using a Qiagen RNEasy Plant Mini kit. First-strand cDNA was synthesized from the extracted RNA using Superscript III First-Strand kit (Invitrogen). The target cDNA was then amplified directly via PCR and sequenced. The CP-specific primers used for the cDNA, PCR and sequencing steps were: forward, 5'-AAGATTG-GCAGGCTA-3'; reverse, 5'-CGGTAAATATTAGAAT-AGCTTG-3'. All sequences generated here have been submitted to GenBank and assigned accession numbers EU371645–EU371650.

A total of six ZYMV CP, newly acquired here, were combined with 49 collected from GenBank (accession numbers available from the authors on request), producing a total dataset of 55 CP sequences, 815 nt in length. To determine the evolutionary relationships among all 55 sequences we employed the maximum-likelihood (ML) method available within the PAUP* package (Swofford, 2003). The best-fit model of nucleotide substitution was determined by MODELTEST (Posada & Crandall, 1998) as TIM+I+Γ4 and this was used as the basis for tree bisection-reconnection branch-swapping (parameter values available from the authors on request). A bootstrap resampling approach (1000 replications), employing the ML substitution model, was used to assess the support for individual nodes. To determine the strength of phylogenetic clustering by country of virus isolation we employed a parsimony character mapping approach (Carrington et al., 2005). Each ZYMV sequence was therefore assigned a character state reflecting its country (or continent) of origin. Given the ML phylogeny for these sequences, the minimum number of state changes needed to produce the observed distribution of country character states was estimated using parsimony (excluding ambiguous changes). To determine the expected number of changes under the null hypothesis of complete mixing among countries, the states of all isolates were randomized 1000 times. The difference between the mean number of observed and expected state changes indicates the level of geographical isolation, with statistical significance assessed by comparing the total number of observed state changes to the number expected under random mixing. All analyses were performed using PAUP* (Swofford, 2003).

The rate of nucleotide substitution per site, as well as the TMRCA of the ZYMV CP sequences were estimated using the Bayesian Markov chain Monte Carlo approach implemented in the BEAST package (Drummond & Rambaut, 2007). This approach analyses the distribution of tip times on millions of plausible sampled phylogenies, so that estimates are set within a rigorous statistical framework. As this analysis requires time-structured data, where the date of sampling of each isolate is known, it was restricted to a subset of 35 CP sequences for which the year of sampling was available, representing a 22 year period from 1984 to 2006. In the case of eight Chinese viruses, sampling dates were known only to the nearest two possible years. To account for this uncertainty, analyses were repeated using the different sampling times available.

We also compared the demographical models of a constant population size and exponential population growth, employing both strict and relaxed (uncorrelated lognormal) molecular clocks. Bayes factors were used to determine the best supported model. Because the TIM+I+Γ4 substitution model is unavailable in the BEAST package, the closely related GTR+I+Γ4 model was used in its place. The extent of statistical uncertainty in parameter estimates is reflected in the 95% highest probability density values. Finally, site-specific selection
pressures in the 55 CP dataset were estimated as the ratio of non-synonymous (dN) to synonymous substitutions (dS) per site (ratio dN/dS) using both the single likelihood ancestor counting (SLAC) and fixed effects likelihood (FEL) methods, available at the Datamonkey facility (Kosakovsky Pond & Frost, 2005).

In accord with other studies of the phylogeography of ZYMV, distinct clusters of viral isolates are apparent in the ML tree of 55 CP sequences (Fig. 1). These clusters represent: (i) a large group of isolates sampled from a variety of locations in Asia (China, Japan, Korea and Taiwan), Europe and the Middle-East (Austria, Germany, Israel, Italy, Hungary and Slovenia), and USA, and previously denoted as groups I and II; (ii) China (previously denoted group III); and (iii) Singapore and the Réunion Island (previously unclassified). We found no compelling evidence for the existence of group II isolates (from Asia), as these fell within the phylogenetic diversity of group I viruses, and suggest that those isolates from Singapore and the Réunion Island are so phylogenetically distinct that they be assigned to their own group.

A number of inferences can be made from this spatial pattern. First, the greatest level of genetic diversity, including the deepest phylogenetic split, is seen in Asia (particularly China), including the presence of one clade of viruses that has only been observed (to date) in China. Although this is compatible with the lineages of ZYMV sampled here having an origin in Asia, this will need to be confirmed with a larger sample of isolates. Second, other than a virus sampled in Florida in 1984, all other USA isolates, sampled between 1992 and 2006 and including those newly obtained from Pennsylvania, have a single common ancestor (Fig. 1). Although the sample size is small, this suggests that there has been some in situ evolution of ZYMV in the USA since this time, without the importation of new viral material. Our parsimony analysis of geographical structure also revealed a strongly significant clustering by country of origin compared with that expected by chance alone (P<0.001). A similarly strong clustering was observed by continent (Americas, Asia, Europe and the Middle-East, Indian Ocean; P<0.001).

Hence, although ZYMV is able to cross geographical boundaries as indicated by the many countries represented within groups I/II, such gene flow is not sufficiently frequent to eradicate geographical structure. More generally, this strong spatial clustering suggests that there is little vertical transmission of ZYMV through cultivated cucurbits, because commercial seeds of cultivated species are likely to be frequently transported across national borders.

The best supported evolutionary model for the CP of ZYMV under our Bayesian coalescent analysis was that of exponential population growth under a relaxed molecular clock (Table 1). Under this model the mean rate of evolutionary change for ZYMV was $5.0 \times 10^{-4}$ nucleotide substitutions per site, per year. Similar rates were obtained under different demographic and molecular clock models, incorporating the different possible sampling times for those viruses for which the exact year of sampling was unknown, and using a range of prior values for the

---

**Fig. 1.** ML tree of 55 ZYMV CP sequences. For viruses where the year of sampling is available, these dates are given in parentheses. Those viruses sampled as part of this study are shaded grey. The group nomenclature depicted represents that previously proposed for ZYMV (Zhao et al., 2003). The tree is drawn to a scale of 0.05 nt substitutions per site and bootstrap values (>90%) are shown next to relevant nodes. The tree is mid-point rooted for clarity only.
substitution rate, indicating that they are robust (results available from the authors on request). This high evolutionary rate falls within the normal range observed in RNA viruses, most of which represent animal RNA viruses (Jenkins et al., 2002; Hanada et al., 2004). As such, we find no evidence that ZYMV evolves any more slowly than animal RNA viruses that are subject to the same, error-prone replication.

Although repeated population bottlenecks undoubtedly influence the genetic structure of viral populations in the short term (Li & Roossinck, 2004), they will have no effect on long term evolutionary rates if most substitutions are selectively neutral. Similarly, although a weaker immune response against plant RNA viruses will reduce the rate at which some non-synonymous mutations accumulate (Garcia-Arenal et al., 2001), the fact that these normally constitute a minor fraction of the total number of nucleotide substitutions means that they are unlikely to have a major impact on long-term evolutionary rates. In support of this we found no evidence for positive selection acting on the CP of ZYMV using either the SLAC or FEL methods; the predominant evolutionary pressure was that of negative (purifying) selection, with a mean $d_{N}/d_{S}$ of 0.108 and 106 of 271 codons negatively selected under the SLAC method. This agrees with previous studies of the CPs of plant RNA viruses, which indicate that they are subject to relatively strong purifying selection (Chare & Holmes, 2004). Further, the lack of positive selection suggests that experimental passage has not had a major impact on our analyses. Although the rapid evolutionary rates observed here for ZYMV will need to be verified for a wider range of plant RNA viruses, the implication from this work is that mutational and replicatory dynamics are similar across a broad range of RNA viruses.

Such high rates of evolutionary change also lead to a recent TMRCA for the isolates of ZYMV analysed here (Table 1). Although there is a relatively large date range because of the inherent sampling error on this analysis (119–771 years), these dates clearly indicate that the spread of this virus has been recent. Indeed, these dates broadly coincide with important ecological changes that may have assisted the spread of ZYMV, including (i) an increase in the number of hectares of worldwide cucurbit cultivation, (ii) the cultivation of cucurbits in novel areas with few wild Cucurbitaceae, facilitating viral transfer from a non-cucurbitaceous plant to the cultivated cucurbits (as observed in a contemporary setting; Perring et al., 1992), and (iii) the cultivation, in close proximity, of cucurbit crops with diverse origins, which allowed the virus to jump to new genera of the family Cucurbitaceae. Overall, our study highlights the utility of gene sequence data to reveal key aspects of the epidemiological history of plant RNA viruses.

### Table 1. Bayesian estimates of population dynamic and evolutionary parameters of the CP gene of ZYMV

<table>
<thead>
<tr>
<th>Parameter estimate</th>
<th>Mean (HPD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date range (years)</td>
<td>22 (1984–2006)</td>
</tr>
<tr>
<td>Sample size</td>
<td>35 sequences</td>
</tr>
<tr>
<td>Substitution rate</td>
<td>$5.0 \times 10^{-4}$ subs per site per year (1.8–8.8 $\times 10^{-4}$ subs per site per year)</td>
</tr>
<tr>
<td>TMRCA</td>
<td>408 years (119–771 years)</td>
</tr>
</tbody>
</table>

References


